Mathematics

QA
47
.H43
v.16
#79

July 1962 · Vol. 16, No. 79

Mathematics of Computation

UNIVERSITY OF MICHIGAN
JUL 16 1962
MATHEMATICS LIBRARY

A journal devoted to advances in numerical analysis, the application of computational methods, mathematical tables, high-speed calculators and other aids to computation

Formerly: Mathematical Tables and other Aids to Computation

Published Quarterly for the
National Academy of Sciences—National Research Council
By the
American Mathematical Society

Editorial Committee

Division of Mathematics

National Academy of Sciences—National Research Council

Washington, D. C.

H. POLACHEK, Chairman, Applied Mathematics Laboratory, David Taylor Model Basin, Washington 7, D. C.
ALAN FLETCHER, University of Liverpool, Liverpool 3, England
P. C HAMMER, University of Wisconsin, Madison 6, Wisconsin
EUGENE ISAACSON, New York University, New York 3, New York
Y. L. LUKE, Midwest Research Institute, Kansas City 10, Missouri
DANIEL SHANKS, Applied Mathematics Laboratory, David Taylor Model Basin, Washington 7, D. C.
A. H. TAUB, University of Illinois, Urbana, Illinois
R. S. VARGA, Case Institute of Technology, Cleveland 6, Ohio
J. W. WRENCH, JR., Applied Mathematics Laboratory, David Taylor Model Basin, Washington 7, D. C.
D. M. YOUNG, Computing Center, University of Texas, Austin 12, Texas

Information to Subscribers

The journal is published quarterly in one volume per year with issues numbered serially since Volume I, Number 1. Starting with January, 1959 subscriptions are $8.00 per year, single copies $2.50. Back issues are available for most issues. Write for specific information.

Volume I (1943–1945), $1.00 per issue.
Volume II (1946–1947), $1.00 per issue.
Volume III (1948–1949), $4.00 per year (four issues), $1.25 per issue.
Volumes IV–XII (1950 through 1958), $5.00 per year, $1.50 per issue.
Volumes XIII–XV (1959–1961), $8.00 per year, $2.50 per issue.

Microcard Edition

Volumes I–X (1943–1956), Nos. 1–56 are now available on Microcards and may be purchased from the Microcard Foundation, Box 2145, Madison 5, Wisconsin, at a cost of $20.00 for the complete set. Succeeding volumes are available on request.

Information to Contributors

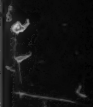
All contributions intended for publication in *Mathematics of Computation* and all books for review should be addressed to H. Polachek, Technical Director, Applied Mathematics Laboratory, David Taylor Model Basin, Washington 7, D. C. The author may suggest an appropriate editor for his paper. Manuscripts should be typewritten double-spaced in the format used by the journal. For journal abbreviations, see *Mathematical Reviews*, v. 21, Index, 1960. Authors should submit the original and one copy, and should retain one copy.

Subscriptions, address changes, business communications and payments should be sent to:

AMERICAN MATHEMATICAL SOCIETY
190 Hope Street
Providence 6, Rhode Island

Published Quarterly for the
NATIONAL ACADEMY OF SCIENCES—NATIONAL RESEARCH COUNCIL
By the
AMERICAN MATHEMATICAL SOCIETY
Providence, Rhode Island
For The

Copyright © 1962, American Mathematical Society
Printed in the United States of America
Second-class postage paid at Providence, Rhode Island and at additional mailing offices.



Some Relations and Values For the Generalized Riemann Zeta Function

By Eldon R. Hansen and Merrell L. Patrick

1. Introduction. In this paper we derive several relations among values of the generalized Riemann Zeta function $\zeta(s, a)$. We show that there is a "fundamental domain" for the variables s and a so that values outside this domain can be obtained from those inside by algebraic operations, using values of the Gamma and trigonometric functions.

We give a table of values of the Zeta function which were calculated using these relations to supplement the formulas commonly used for such calculations. Table 1 gives $\zeta(s, a)$ to seventeen decimal places for $a = \frac{1}{4}$ and $\frac{3}{4}$ and for $s = -6(\frac{1}{3})6\frac{1}{3}$. The ordinary Riemann Zeta function is also given to seventeen decimal places for $s = 0(\frac{1}{3})6\frac{1}{3}$ in Table 2. These tables are useful in diffraction theory [10], [11].

2. Relations among Zeta Functions. Consider the defining relation

$$\zeta(s, a) = \sum_{n=0}^{\infty} \frac{1}{(n + a)^s} \tag{1}$$

which holds for $\mathrm{Re}(s) > 1$ and all values of a except zero or negative integers. (One usually assumes s complex but a real.) We first note that

$$\begin{aligned} \zeta(s, a + 1) &= \sum_{n=0}^{\infty} \frac{1}{(n + a + 1)^s} \\ &= \sum_{n=0}^{\infty} \frac{1}{(n + a)^s} - \frac{1}{a^s} \\ &= \zeta(s, a) - a^{-s}. \end{aligned} \tag{2}$$

By analytic continuation, this relation holds for all s. Thus if $\zeta(s, a)$ is known for all s, but only for values of a satisfying $a_0 < a \leqq a_0 + 1$ for some a_0, then we can calculate $\zeta(s, a)$ for any value of a by algebraic operations. One usually specifies $a_0 = 0$, so that the "fundamental interval" is $0 < a \leqq 1$.

We now show that this fundamental interval in a need only be of length $\frac{1}{2}$. In the defining relation (1), let $a = (1/q) - b$ where q is an integer. Then

$$\begin{aligned} \zeta\left(s, \frac{1}{q} - b\right) &= q^s \sum_{n=0}^{\infty} \frac{1}{(nq + 1 - bq)^s} \\ &= q^s \left[\sum_{n=0}^{\infty} \frac{1}{(n + 1 - bq)^s} - \sum_{r=2}^{q} \sum_{n=0}^{\infty} \frac{1}{(nq + r - bq)^s}\right] \\ &= q^s \zeta(s, 1 - bq) - \sum_{r=2}^{q} \zeta\left(s, \frac{r}{q} - b\right). \end{aligned}$$

Received January 10, 1962.

THE UNIVERSITY OF MICHIGAN LIBRARIES

Therefore

$$\sum_{r=1}^{q} \zeta\left(s, \frac{r}{q} - b\right) = q^s \zeta(s, 1 - bq). \tag{3}$$

By analytic continuation, this result holds for all s. This relation was obtained by Powell [7] for the special case $s = \frac{1}{2}$.

Equation (3) yields many interesting and useful results. For example, let $b = \frac{1}{2} - c$. Then, for $q = 2$, we obtain

$$\zeta(s, c) = 2^s \zeta(s, 2c) - \zeta(s, c + \tfrac{1}{2}). \tag{4}$$

For $c = \frac{1}{2}$ this yields the well known result

$$\zeta(s, \tfrac{1}{2}) = (2^s - 1)\zeta(s), \tag{5}$$

where $\zeta(s)$ is the ordinary Riemann Zeta function defined by $\zeta(s, 1) = \zeta(s)$.

Now assume that $\zeta(s, a)$ is known for all s and for all a in the interval $\frac{1}{2} < a \leqq 1$. Then equation (4) shows that, if a_1 is such that $0 < a_1 \leqq \frac{1}{2}$, then $\zeta(s, a_1)$ can be expressed in terms of values assumed known. We merely apply relation (4) repeatedly as needed. An inductive proof can be easily obtained but will not be given here. As an example of the use of equation (4), consider the case $a_1 = \frac{1}{8}$. From equation (4),

$$\zeta(s, \tfrac{1}{8}) = 2^s \zeta(s, \tfrac{1}{4}) - \zeta(s, \tfrac{5}{8}). \tag{6}$$

The argument $\frac{1}{4}$ is not in the desired interval, so we use (4) to obtain

$$\zeta(s, \tfrac{1}{4}) = 2^s \zeta(s, \tfrac{1}{2}) - \zeta(s, \tfrac{3}{4}). \tag{7}$$

Again, we do not know $\zeta(s, \frac{1}{2})$, so we use (4) to obtain

$$\zeta(s, \tfrac{1}{2}) = 2^s \zeta(s, 1) - \zeta(s, 1) = (2^s - 1)\zeta(s). \tag{8}$$

Combining (6), (7) and (8), we have

$$\zeta(s, \tfrac{1}{8}) = 2^{2s}(2^s - 1)\zeta(s) - 2^s \zeta(s, \tfrac{3}{4}) - \zeta(s, \tfrac{5}{8}).$$

All the functions in the right member are assumed known, hence $\zeta(s, \frac{1}{8})$ is obtained by algebraic operations.

For special values of a, the more general relation (3) may be used more fruitfully than the special form (4).

We now see that $\frac{1}{2} < a \leqq 1$ is a fundamental interval for $\zeta(s, a)$. We next show that for rational values of a, we can obtain $\zeta(s, a)$ for $\mathrm{Re}(s) < \frac{1}{2}$ from values for $\mathrm{Re}(s) \geqq \frac{1}{2}$.

Consider the Hurwitz formula (see [1], page 269)

$$\frac{(2\pi)^s}{2\Gamma(s)} \zeta(1 - s, a) = \cos\frac{\pi s}{2} \sum_{n=1}^{\infty} \frac{\cos 2\pi n a}{n^s} + \sin\frac{\pi s}{2} \sum_{n=1}^{\infty} \frac{\sin 2\pi n a}{n^s}, \tag{9}$$

which is valid for $\mathrm{Re}(s) > 1$ and $0 < a \leqq 1$. Let $a = p/q$, where p and q are in-

tegers (which need not be relatively prime). Then

$$\frac{(2\pi)^s}{2\Gamma(s)}\zeta\left(1-s,\frac{p}{q}\right)=\cos\frac{\pi s}{2}\sum_{n=1}^{\infty}\frac{\cos 2\pi np/q}{n^s}+\sin\frac{\pi s}{2}\sum_{n=1}^{\infty}\frac{\sin 2\pi np/q}{n^s}$$

$$=\cos\frac{\pi s}{2}\sum_{r=0}^{q-1}\sum_{n=1}^{\infty}\frac{\cos 2\pi(qn-r)p/q}{(qn-r)^s}+\sin\frac{\pi s}{2}\sum_{r=0}^{q-1}\sum_{n=1}^{\infty}\frac{\sin 2\pi(qn-r)p/q}{(qn-r)^s}$$

$$=\frac{1}{q^s}\cos\frac{\pi s}{2}\sum_{r=0}^{q-1}\sum_{n=1}^{\infty}\frac{\cos 2\pi rp/q}{(n-r/q)^s}-\frac{1}{q^s}\sin\frac{\pi s}{2}\sum_{r=0}^{q-1}\sum_{n=1}^{\infty}\frac{\sin 2\pi rp/q}{(n-r/q)^s}$$

$$=\frac{1}{q^s}\sum_{r=0}^{q-1}\left[\cos\frac{\pi s}{2}\cos\frac{2\pi rp}{q}-\sin\frac{\pi s}{2}\sin\frac{2\pi rp}{q}\right]\sum_{n=0}^{\infty}\frac{1}{(n+1-r/q)^s}.$$

Replacing r by $q-r$ and simplifying,

$$\zeta(1-s,p/q)=\frac{2\Gamma(s)}{(2\pi q)^s}\sum_{r=1}^{q}\cos\pi\left(\frac{s}{2}-\frac{2rp}{q}\right)\zeta(s,r/q). \tag{10}$$

By analytic continuation, this result holds for all s. Thus, if $\zeta(s,a)$ is known for $\mathrm{Re}(s)\geqq\frac{1}{2}$ and all rational a, we can compute $\zeta(s,a)$ for all s and all rational a, $0<a\leqq 1$, from equation (10). Conversely, if we know $\zeta(s,a)$ for all rational a, $0<a\leqq 1$, and for $\mathrm{Re}(s)\leqq\frac{1}{2}$, we can compute $\zeta(s,a)$ from equation (10) for; all s except for integer values. The right hand side of equation (10) contains $\Gamma(s)$, which is infinite for s a negative integer. Hence equation (10) cannot, in general, be used directly to yield $\zeta(s,a)$ for s a positive integer. However equations (2), (4), and (10) enable us to find $\zeta(s,a)$ for all s and all rational a from values of $\zeta(s,a)$ for $\mathrm{Re}(s)\geqq\frac{1}{2}$ and $\frac{1}{2}<a\leqq 1$.

Let us now consider some special cases of equation (10). For $p=1$ and $q=2$, we again obtain equation (5). For $p=q=1$, we obtain the Riemann relation

$$\zeta(1-s)=\frac{2\Gamma(s)}{(2\pi)^s}\cos\frac{\pi s}{2}\zeta(s). \tag{11}$$

For the general case $p=q$, we obtain, by using relation (11),

$$\sum_{r=1}^{q-1}\zeta(s,r/q)=(q^s-1)\zeta(s) \tag{12}$$

which also follows from equation (3). If we also write equation (12) with s replaced by $1-s$, then the resulting equation can be combined with equations (11) and (12) to yield

$$\pi^{-s/2}\Gamma(s/2)(q^s-1)\sum_{r=1}^{q-1}\zeta(s,r/q)$$
$$=\pi^{-(1-s)/2}\Gamma\left(\frac{1-s}{2}\right)(q^{1-s}-1)\sum_{r=1}^{q-1}\zeta(1-s,r/q). \tag{13}$$

That is, the members of this equation are invariant under replacement of s by $1-s$. Hence we may regard equation (13) as a generalization of the Riemann

THE UNIVERSITY OF MICHIGAN LIBRARIES

relation (11), which is often written as

$$\Gamma\left(\frac{1-s}{2}\right)\pi^{-(1-s)/2}\zeta(1-s) = \Gamma\left(\frac{s}{2}\right)\pi^{-s/2}\zeta(s).$$

Next substitute $q = 4$ and $p = 1$ and 3 in equation (10). In this case, we obtain, using equations (5) and (11)

$$\zeta(1-s, \tfrac{1}{2} \pm \tfrac{1}{4}) = \frac{2-2^s}{4^s}\zeta(1-s) \pm \frac{2\Gamma(s)}{(8\pi)^s}\sin\frac{\pi s}{2}[\zeta(s, \tfrac{3}{4}) - \zeta(s, \tfrac{1}{4})]. \tag{14}$$

Subtracting this equation, using the lower sign from the equation using the upper sign, we obtain

$$\zeta(1-s, \tfrac{3}{4}) - \zeta(1-s, \tfrac{1}{4}) = \frac{4\Gamma(s)}{(8\pi)^s}\sin\frac{\pi s}{2}[\zeta(s, \tfrac{3}{4}) - \zeta(s, \tfrac{1}{4})]. \tag{15}$$

Now the Dirichlet L-function is defined to be

$$L(s) = \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)^s}$$

for $\mathrm{Re}(s) > 1$. This can be expressed as

$$L(s) = 4^{-s}[\zeta(s, \tfrac{1}{4}) - \zeta(s, \tfrac{3}{4})].$$

Hence equation (15) can be rewritten as

$$L(1-s) = (2/\pi)^s\Gamma(s)\sin\frac{\pi s}{2}L(s).$$

This relation is derived (but misprinted) by Titchmarsh (see [6], page 66).

For m equal to a positive integer,

$$\zeta(-m, a) = -\frac{B_{m+1}(a)}{(m+1)}, \tag{16}$$

where $B_m(a)$ is the mth Bernoulli polynomial in a. (See [1], page 267.) Hence, letting $s \to -2m$ in equation (14), and using the fact that $B_m(1-a) = (-1)^m B_m(a)$, we obtain

$$\zeta(2m+1, \tfrac{1}{2} \pm \tfrac{1}{4}) = 2^{2m}(2^{2m+1}-1)\zeta(2m+1) \pm \frac{2^{6m+1}\pi^{2m+1}}{(2m+1)!}B_{2m+1}(\tfrac{1}{4}).$$

From these equations, we find that

$$L(2m+1) = -\frac{(2\pi)^{2m+1}}{2(2m+1)!}B_{2m+1}(\tfrac{1}{4}).$$

Many other interesting results can be obtained by looking at special cases of equation (10).

The method used to derive equations (3) and (10) can also be applied to the more general function

$$\zeta(s, a, z) = \sum_{n=0}^{\infty}\frac{z^n}{(n+a)^s}$$

discussed by Mitchell [8], Lerch [9] and others.

3. Calculation of the Table. In the discussion above we showed, by using equation (10), that we could find $\zeta(s, a)$ for the $\mathrm{Re}(s) \leqq \frac{1}{2}$ and rational a, $0 < a \leqq 1$, provided we know $\zeta(s, a)$ for $\mathrm{Re}(s) > \frac{1}{2}$ and rational a, $0 < a \leqq 1$. The equation

$$\zeta(s, a) = \sum_{n=0}^{N-1} \frac{1}{(n+a)^s} - \frac{(N+a)^{1-s}}{1-s} + \frac{(N+a)^{-s}}{2} - \sum_{n=1}^{\infty} \frac{B_n}{(2n)!} \frac{\Gamma(1-s)}{\Gamma(2-s-2n)} (N+a)^{-s-(2n-1)}, \tag{17}$$

where $B_1 = \frac{1}{6}$, $B_2 = -\frac{1}{30}$, $B_3 = \frac{1}{42}$, $B_4 = -\frac{1}{30}$, $B_5 = \frac{5}{66}$, $\cdots$ are Bernoulli numbers, is a representation of $\zeta(s, a)$ for $\mathrm{Re}(s) > 0$, $s \neq 1$, and $0 < a \leqq 1$. We use this asymptotic series to calculate $\zeta(s, \frac{1}{4})$, $\zeta(s, \frac{3}{4})$, and $\zeta(s, 1) = \zeta(s)$ for $s = \frac{2}{3}(\frac{1}{2})^{\frac{2}{3}}$. It is not necessary to use this expression to calculate $\zeta(s, \frac{1}{2})$ because of equation (8). With these values of $\zeta(s, \frac{1}{4})$, $\zeta(s, \frac{3}{4})$, and $\zeta(s, 1)$ for $s = \frac{2}{3}(\frac{1}{2})^{\frac{2}{3}}$, we calculate $\zeta(s, \frac{1}{4})$ and $\zeta(s, \frac{3}{4})$, using (14), for $s = -\frac{6}{5}(\frac{1}{2})^{\frac{1}{2}}$ with the exception $s = 0$. For $s = 0$ we use the identity $\zeta(0, a) = \frac{1}{2} - a$. The Gamma function, $\Gamma(s)$, appearing on the right side of equation (10) is calculated by making use of the asymptotic series (see [1] Section 12.33)

$$\log \Gamma(s) = (s - \tfrac{1}{2}) \log s - s + \tfrac{1}{2} \log (2\pi) + \sum_{r=1}^{\infty} \frac{(-1)^{r-1} B_r}{2r(2r-1)s^{2r-1}}$$

where B_r are the Bernoulli numbers. This expression holds for $|\arg s| \leqq \frac{1}{2}\pi - \Delta$ and $0 < \Delta < \frac{1}{4}\pi$. Using this relation in conjunction with the equation $\Gamma(s+1) = s\Gamma(s)$, we obtain the desired values of the Gamma function.

Equation (17), for $a = 1$, is equivalent to the expression used by Gram [2] and by Haselgrove [3] in preparing their tables of the ordinary Riemann Zeta function.

4. Methods of Checking the Table Values. Equation (16) was used to check the table entries for negative integer values of s, while Hurwitz's formula (9) was used to check the entries for $s = -\frac{6}{5}(\frac{1}{2}) - \frac{2}{5}$. For positive integer values of s, the values of $\zeta(s)$ agree exactly with the table in [5], page xxv. Our table was not checked for other values of s, since no other simple method seemed to present itself. Note, however, the table for negative values of s was calculated using the entries for positive values of s. Hence, it seems safe to assume that the table entries for positive values of s are correct to at least as many digits as the table entries for corresponding negative values of s.

5. Comments Regarding the Calculation of the Table. The table shown below was calculated using a precision of seventy binary digits, which is approximately equivalent to twenty-one decimal digits.

It was impossible to enter most of the values of s exactly in the computer, since 1/3 does not have a finite binary representation. Therefore, we should consider the errors introduced in our table from using a truncated binary representation. We were able to show, by using a truncated Taylor series, that the errors so introduced did not affect the accuracy of the table. The various calculations involved in showing the insignificance of these errors were not included in this paper, since they were quite straightforward.

UNIVERSITY OF MICHIGAN LIBRARIES

6. The Table of Values. The tabular data have been listed to seventeen significant decimal digits. The decimal point is located immediately to the left of the left-most digit of each entry, while the two-digit integer at the right denotes the exponent of the power of ten by which the entry is to be multiplied.

Lockheed Missiles and Space Company
Palo Alto, California

1. E. T. Whittaker & G. N. Watson, *A Course in Modern Analysis*, fourth edition, Cambridge, 1952.
2. J. P. Gram, "Tafeln für die Riemannsche Zetafunktion," *Kungl. Danske* Vid. Selsk. Skr. (8), v. 10, 1925, p. 313–325.
3. C. B. Haselgrove, *Tables of the Riemann Zeta Function*, Cambridge University Press, 1960.
4. R. Hensman, *Tables of the Generalized Riemann Zeta Function*, Report No. T 2111, Telecommunications Research Establishment, Ministry of Supply, Great Malvern, Worcestershire, 1948.
5. British Association for the Advancement of Science, Mathematical Tables, Vol. I, *Circular and Hyperbolic Functions*, third edition, Cambridge University Press, 1951.
6. E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, second edition, Oxford University Press, 1948.
7. E. O. Powell, "A table of the generalized Riemann zeta function in a particular case," *Quart. J. Mech. Appl. Math.* v. 5, 1952, p. 116–119.
8. K. Mitchell, "Tables of a function and related functions," *Phil. Mag.*, v. 40, 1949, p. 351–368.
9. E. Lerch, "Note sur la fonction $R(w, x, s) = \sum_0^\infty \frac{e^{2k\pi ix}}{(w+k)^s}$," *Acta. Math.* (Stockholm), v. 11, 1887, p. 19–24.
10. H. Bremmer, *Terrestrial Radio Waves*, Elsevier Publishing Co., New York, 1949.
11. Nelson Logan, *General Research in Diffraction Theory*, v. I., Lockheed Missiles and Space Division Report #288087, December 1959.

TABLE 1

s	$\zeta(s, \frac{1}{4})$	$\zeta(s, \frac{3}{4})$
64/3	69814 63658 33156 77 13	46276 41642 09957 54 03
21	43980 46511 10400 92 13	42044 91481 42985 79 03
62/3	27705 95688 87832 77 13	38200 34049 99724 72 03
61/3	17453 65914 58290 05 13	34707 31296 92456 16 03
20	10995 11627 77601 15 13	31533 68689 82531 57 03
59/3	69264 89222 19591 87 12	28650 25632 40314 70 03
58/3	43634 14786 45735 82 12	26030 48587 06422 96 03
19	27487 79069 44014 41 12	23650 26655 19854 06 03
56/3	17316 22305 54910 39 12	21487 69390 40321 69 03
55/3	10908 53696 61447 33 12	19522 86640 47054 77 03
18	68719 47673 60180 15 11	17737 70232 63351 39 03
53/3	43290 55763 87431 22 11	16115 77333 52502 14 03
52/3	27271 34241 53785 57 11	14642 15330 71921 73 03
17	17179 86918 40225 19 11	13303 28096 72635 09 03
50/3	10822 63940 97051 87 11	12086 83509 03545 36 03
49/3	68178 35603 86554 48 10	10981 62111 35873 48 03
16	42949 67296 02814 98 10	99774 68117 34460 20 02
47/3	27056 59852 45055 63 10	90651 35226 87845 72 02
46/3	17044 58900 99251 96 10	82362 26573 20047 47 02
15	10737 41824 03518 96 10	74831 14030 73469 44 02
44/3	67641 49631 42966 14 09	67988 67020 79914 81 02
43/3	42611 47252 80800 17 09	61771 88735 02511 35 02
14	26843 54560 43992 27 09	56123 58191 92793 72 02
41/3	16910 37408 23656 51 09	50991 77593 55328 03 02
40/3	10652 86813 61045 94 09	46329 24497 93384 86 02
13	67108 86405 50022 12 08	42093 08367 45511 12 02
38/3	42275 93525 33217 14 08	38244 31093 55480 72 02
37/3	26632 17039 13370 57 08	34747 51134 84446 15 02
12	16777 21606 87796 31 08	31570 50939 08321 59 02
35/3	10568 98387 26213 03 08	28684 07349 76745 21 02
34/3	66580 42661 71997 90 07	26061 64725 71060 80 02
11	41943 04086 03548 00 07	23679 10527 14066 46 02
32/3	26422 46042 34022 87 07	21514 53144 78372 37 02
31/3	16645 10745 38096 69 07	19548 01769 36192 33 02
10	10485 76107 68311 48 07	17761 48118 13370 04 02
29/3	66056 16034 78926 59 06	16138 49852 76111 40 02
28/3	41612 77864 98520 10 06	14664 15539 12355 32 02
9	26214 41349 21724 08 06	13324 91014 94593 48 02
26/3	16514 05173 61085 33 06	12108 47045 54080 94 02
25/3	10403 20723 10878 44 06	11003 68161 90407 28 02
8	65536 16938 67090 17 05	10000 42589 27890 58 02
23/3	41285 27582 34437 32 05	90895 31887 04304 92 01
22/3	26008 17634 03920 72 05	82626 93498 64416 39 01
7	16384 21345 51995 72 05	75123 97920 96579 80 01
20/3	10321 50411 38500 09 05	68318 62538 56294 03 01
19/3	65022 44664 70976 70 04	62149 80828 62956 03 01
6	40962 70948 06401 08 04	56562 77857 28816 43 01
17/3	25806 12469 70992 37 04	51508 76680 73079 38 01
16/3	16258 18668 37913 66 04	46944 78141 49920 25 01
5	10243 48974 52658 06 04	42833 58575 64239 47 01
14/3	64546 14107 47051 67 03	39143 93590 59400 43 01
13/3	40679 42509 60218 28 03	35851 22960 35115 28 01

THE UNIVERSITY OF MICHIGAN LIBRARIES

TABLE 1—*Continued*

s	$\zeta(s, \frac{1}{4})$		$\zeta(s, \frac{3}{4})$	
4	25646 36906 68198 07	03	32938 85422 47510 00	01
11/3	16178 62720 13961 71	03	30400 81320 64265 08	01
10/3	10217 48139 83634 30	03	28246 87653 98801 74	01
3	64663 86996 87684 60	02	26513 16608 16881 98	01
8/3	41092 64168 75447 99	02	25285 74853 60394 04	01
7/3	26334 77863 02380 64	02	24759 96055 02275 82	01
2	17197 32915 45071 11	02	25418 79647 67160 65	01
5/3	11787 30284 47967 72	02	28747 02273 10830 06	01
4/3	95676 33344 56813 25	01	42231 01605 70504 20	01
1	∞		∞	
2/3	−24411 86144 96889 78	00	−20381 06131 82465 56	01
1/3	33101 39009 27282 69	00	−64976 99170 75290 43	00
0	25000 00000 00000 00	00	−25000 00000 00000 00	00
−1/3	13499 19957 79665 53	00	−89579 84483 88113 98	−01
−2/3	55399 96817 77458 36	−01	−19221 97815 54288 52	−01
−1	10416 66666 66666 67	−01	10416 66666 66666 67	−01
−4/3	−10558 77278 36124 00	−01	20147 85800 02319 75	−01
−5/3	−16992 51822 29812 52	−01	20093 86425 54835 96	−01
−2	−15625 00000 00000 00	−01	15625 00000 00000 00	−01
−7/3	−10834 45058 95508 42	−01	97268 16010 89067 10	−02
−8/3	−52815 71878 70004 44	−02	40704 00253 32719 07	−02
−3	−45572 91666 66666 67	−03	−45572 91666 66666 67	−03
−10/3	29301 12394 80509 84	−02	−34597 28641 05474 40	−02
−11/3	46698 19500 95810 35	−02	−48805 58903 88315 47	−02
−4	48828 12500 00000 00	−02	−48828 12500 00000 00	−02
−13/3	38899 36684 63823 16	−02	−37822 44498 22381 09	−02
−14/3	21239 95040 97140 41	−02	−19861 15110 11172 00	−02
−5	60066 34424 60317 46	−04	60066 34424 60317 46	−04
−16/3	−18432 75029 83014 06	−02	19233 18930 79122 45	−02
−17/3	−31960 86810 98371 30	−02	32323 03687 52690 05	−02
−6	−37231 44531 25000 00	−02	37231 44531 24500 00	−02
−19/3	−33017 11705 70240 87	−02	32782 87245 91040 61	−02
−20/3	−19865 27145 97911 51	−02	19530 95592 98213 26	−02
−7	−16148 88509 11458 33	−04	−16148 88509 11458 33	−04
−22/3	22055 37817 57697 42	−02	−22292 83934 16781 00	−02
−23/3	41577 48345 17227 95	−02	−41695 51768 48320 96	−02
−8	52833 55712 89062 50	−02	−52833 55712 89062 50	−02
−25/3	50987 50282 80912 53	−02	−50896 45120 98702 91	−02
−26/3	33229 98874 70615 11	−02	−33088 82387 94668 38	−02
−9	73837 51146 72111 74	−05	73837 51146 72111 74	−05
−28/3	−43631 40309 49183 10	−02	43748 62335 98758 29	−02
−29/3	−88331 79840 49455 97	−02	88394 53105 86276 23	−02
−10	−12045 14503 47900 39	−01	12045 14503 47900 39	−01
−31/3	−12448 12326 45933 15	−01	12442 55605 16962 73	−01
−32/3	−86654 89190 70478 86	−02	86562 63379 78600 91	−02
−11	−51470 93963 85593 44	−05	−51470 93963 85593 44	−05
−34/3	−12956 24917 98495 48	−01	−12964 94737 89903 23	−01
−35/3	27851 69352 45565 33	−01	−27856 63944 24212 63	−01
−12	40274 33693 40896 61	−01	−40274 33693 40896 61	−01
−37/3	44066 74225 89134 15	−01	−44061 81300 66818 43	−01
−38/3	32424 26287 38385 85	−01	−32415 62726 12390 43	−01
−13	50856 42139 11692 30	−05	50856 42139 11692 30	−05
−40/3	−53981 95160 35609 58	−01	53991 01083 71719 25	−01

TABLE 1—*Continued*

s	$\zeta(s, \frac{1}{4})$	$\zeta(s, \frac{3}{4})$
−41/3	−12213 66720 37044 43 00	12214 20945 65951 94 00
−14	−18566 93821 02787 49 00	18566 93821 02787 49 00
−43/3	−21331 38708 22892 45 00	21330 79048 86111 17 00
−44/3	−16461 18663 02940 22 00	16460 09042 37713 87 00
−15	−67634 01438 99342 01 −05	−67634 01438 99342 01 −05
−46/3	30053 84266 68572 82 00	−30055 10353 95330 80 00
−47/3	71088 87471 07870 00 00	−71089 66375 73979 33 00
−16	11287 34563 53554 50 01	−11287 34563 53554 50 01
−49/3	13532 13201 24995 85 01	−13532 03739 35347 35 01
−50/3	10887 19301 23970 26 01	−10887 01175 14453 29 01
−17	11649 82235 12560 03 −04	11649 82235 12560 03 −04
−52/3	−21552 81617 38443 47 01	21553 04222 79286 09 01
−53/3	−53020 89805 52583 77 01	53021 04518 12564 35 01
−18	−87489 01292 85995 50 01	87489 01292 85995 50 01
−55/3	−10892 53279 53442 03 02	10892 51375 45982 56 02
−56/3	−90944 16979 46386 51 01	90943 79125 83854 77 01
−19	−25230 56188 58262 09 −04	−25230 56188 58262 09 −04
−58/3	19350 58818 71838 18 02	−19350 63892 61704 53 02
−59/3	49304 35983 54058 91 02	−49304 39403 87055 58 02
−20	84212 65834 78432 17 02	−84212 65834 78432 17 02
−61/3	10846 36397 60161 63 03	−10846 35923 59976 00 03

TABLE 2

s	$\zeta(s)$	s	$\zeta(s)$
64/3	10000 00378 53233 25 01	31/3	10007 87512 59735 22 01
21	10000 00476 93298 68 01	10	10009 94575 12781 81 01
62/3	10000 00600 91541 53 01	29/3	10012 56544 89496 33 01
61/3	10000 00757 13123 99 01	28/3	10015 88198 62306 52 01
20	10000 00953 96203 39 01	9	10020 08392 82608 22 01
59/3	10000 01201 96944 13 01	26/3	10025 41240 15797 39 01
58/3	10000 01514 46249 87 01	25/3	10032 17649 68406 42 01
19	10000 01908 21271 66 01	8	10040 77356 19794 43 01
56/3	10000 02404 35545 91 01	23/3	10051 71613 23892 26 01
55/3	10000 03029 52623 90 01	22/3	10065 66797 08110 32 01
18	10000 03817 29326 50 01	7	10083 49277 38192 28 01
53/3	10000 04809 96364 94 01	20/3	10106 32075 32048 38 01
52/3	10000 06060 86098 24 01	19/3	10135 64087 21870 02 01
17	10000 07637 19763 79 01	6	10173 43061 98444 91 01
50/3	10000 09623 69761 94 01	17/3	10222 34193 45949 71 01
49/3	10000 12127 16680 42 01	16/3	10285 97325 52868 25 01
16	10000 15282 25940 87 01	5	10369 27755 14336 99 01
47/3	10000 19258 75533 22 01	14/3	10479 19229 28351 53 01
46/3	10000 24270 74650 59 01	13/3	10625 64589 86109 19 01
15	10000 30588 23630 70 01	4	10823 23233 71113 82 01
44/3	10000 38551 79067 62 01	11/3	11094 13692 59748 37 01
43/3	10000 48591 05077 17 01	10/3	11473 56236 88274 29 01
14	10000 61248 13505 87 01	3	12020 56903 15959 43 01
41/3	10000 77207 23697 67 01	8/3	12841 90540 23974 21 01
40/3	10000 97332 08000 25 01	7/3	14151 55609 44598 30 01
13	10001 22713 34757 85 01	2	16449 34066 84822 64 01
38/3	10001 54728 79053 14 01	5/3	21235 22968 85758 35 01
37/3	10001 95119 46821 60 01	4/3	36009 37750 45886 24 01
12	10002 46086 55330 80 01	1	∞
35/3	10003 10414 39287 00 01	2/3	−24475 80736 23365 82 01
34/3	10003 91627 16273 87 01	1/3	−97336 02483 50782 72 01
11	10004 94188 60411 95 01	0	−50000 00000 00000 00 00
32/3	10006 23757 16251 28 01		

Euler's Constant to 1271 Places

By Donald E. Knuth

Abstract. The value of Euler's or Mascheroni's constant

$$\gamma = \lim_{n\to\infty} (1 + \tfrac{1}{2} + \cdots + (1/n) - \ln n)$$

has now been determined to 1271 decimal places, thus extending the previously known value of 328 places. A calculation of partial quotients and best rational approximations to γ was also made.

1. Historical Background. Euler's constant was, naturally enough, first evaluated by Leonhard Euler, and he obtained the value 0.577218 in 1735 [1]. By 1781 he had calculated it more accurately as 0.5772156649015325 [2]. The calculations were carried out more precisely by several later mathematicians, among them Gauss, who obtained

$$\gamma = 0.57721566490153286060653.$$

Various British mathematicians continued the effort [3], [4]; an excellent account of the work done on evaluation of γ before 1870 is given by Glaisher [5]. Finally, the famous mathematician-astronomer J. C. Adams [6] laboriously determined γ to 263 places. Adams thereby extended the work of Shanks, who had obtained 110 places (101 of which were correct).

Adams' result stood until 1952, when Wrench [9] calculated 328 decimal places. Although much work has been done trying to decide whether γ is rational, the evaluation has not been carried out any more precisely. With the use of high-speed computers, the constants π and e have been evaluated to many thousands of decimal places [11], [12]. A complete bibliography for π appears in [11]. The evaluation of γ to many places is considerably more difficult.

2. Evaluation of γ. The technique used here to calculate γ is essentially that used by Adams and earlier mathematicians. A complete derivation of the method is given by Knopp [7]. We use Euler's summation formula in the form

$$\sum_{i=1}^{n} f(i) = \int_1^n f(x)\,dx + \tfrac{1}{2}(f(n) + f(1)) + \sum_{j=1}^{k} \frac{B_{2j}}{(2j)!}[f^{(2j-1)}(n) - f^{(2j-1)}(1)] + R_k \tag{1}$$

where B_m are the Bernoulli numbers defined symbolically by

$$e^{Bx} = \frac{x}{e^x - 1}. \tag{2}$$

With this notation, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$, $B_4 = -\frac{1}{30}$, etc. Here the remainder

Received January 12, 1962.

R_k is given by

$$R_k = \frac{1}{(2k+1)!}\int_1^n P_{2k+1}(x) f^{(2k+1)}(x)\,dx; \tag{3}$$

and $P_{2k+1}(x)$ is a periodic Bernoulli polynomial, symbolically

$$P_{2k+1}(x) = (\{x\} + B)^{2k+1} = (-1)^{k-1}(2k+1)!\sum_{r=1}^{\infty}\frac{2\sin 2r\pi x}{(2r\pi)^{2k+1}} \tag{4}$$

where $\{x\}$ is the fractional part of x.

Now we put $f(x) = 1/x$, obtaining from (1)

$$\begin{aligned} 1 + \tfrac{1}{2} + \cdots + \frac{1}{n} &= \ln n + \frac{1}{2} + \frac{1}{2n} \\ &\quad + \frac{B_2}{2}\left(1 - \frac{1}{n^2}\right) + \cdots + \frac{B_{2k}}{2k}\left(1 - \frac{1}{n^{2k}}\right) - \int_1^n \frac{P_{2k+1}(x)}{x^{2k+2}}\,dx. \end{aligned} \tag{5}$$

Taking the limit in (5) as $n \to \infty$, we find

$$\gamma = \frac{1}{2} + \frac{B_2}{2} + \cdots + \frac{B_{2k}}{2k} - \int_1^{\infty}\frac{P_{2k+1}(x)}{x^{2k+2}}\,dx. \tag{6}$$

Subtracting (5) from (6) gives

$$\begin{aligned} \gamma = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n - \frac{1}{2n} + \frac{B_2}{2n^2} \\ + \cdots + \frac{B_{2k}}{(2k)n^{2k}} - \int_n^{\infty}\frac{P_{2k+1}(x)\,dx}{x^{2k+2}}. \end{aligned} \tag{7}$$

If the remainder is discarded and we consider (7) as an infinite series in k, it diverges as $k \to \infty$. It still yields a good method for calculating γ, however, since

$$|P_{2k+1}(x)| \leqq \frac{2(2k+1)!}{(2\pi)^{2k+1}}\sum_{r=1}^{\infty}\frac{1}{r^{2k+1}}, \tag{8}$$

and by applying Stirling's formula to (8) we obtain

$$\left|\int_n^{\infty}\frac{P_{2k+1}(x)\,dx}{x^{2k+2}}\right| \leqq \frac{4}{n}\sqrt{\frac{k}{\pi}}\left(\frac{k}{n\pi e}\right)^{2k}. \tag{9}$$

Put $k = 250$ and $n = 10000$ to obtain a remainder

$$\left|\int_{10000}^{\infty}\frac{P_{501}(x)\,dx}{x^{502}}\right| < 10^{-1269}, \tag{10}$$

so these values may be used in (7) to determine γ to at least 1269 places. This particular choice of k and n was made for convenience on a decimal computer, in an attempt to obtain the greatest precision in a reasonable time.

3. Details of the Computation. The sum $1 + \frac{1}{2} + \cdots + \frac{1}{10000}$ was evaluated as

$$S_{10000} = \frac{3}{2} + \frac{7}{12} + \cdots + \frac{19999}{99990000} = 9.787606036\cdots. \tag{11}$$

Combining terms in this way reduced the number of necessary divisions. The natural

logarithm of 10000 was then determined by

$$\ln 10000 = -252 \ln (1 - .028) + 200 \ln (1 + .0125) + 92 \ln (1 - .004672). \tag{12}$$

Such an expansion was designed for fast convergence and for convenience on a decimal computer. It is a simple matter to obtain such an expansion by hand calculation; we seek integers (x, y, z) such that $2^x 3^y 5^z \approx 1$ and $y \geqq 0$. If three linearly independent solutions are obtained, one can calculate ln 2, ln 3, and ln 5, and, in particular, ln 10. If $2^{x_1}3^{y_1}5^{z_1} > 1$ and $2^{x_2}3^{y_2}5^{z_2} < 1$, suitable positive integral combinations of (x_1, y_1, z_1) and (x_2, y_2, z_2) will give closer approximations. The method is to find small values of (x, y, z) so that $x + y \log_2 3 + z \log_2 5 \approx 0$, then combine these to get better and better approximations. The expansion (12) corresponds to the solutions $(-1, 5, -3)$, $(-4, 4, -1)$, and $(6, 5, -6)$. For a binary computer the extra requirement $z \geqq 0$ makes it more difficult, but solutions can be used such as

$$\ln 10000 = 160 \ln 2^{-32}3^{7}5^{9} - 864 \ln 2^{-11}3^{4}5^{2} + 292 \ln 2^{-15}3^{8}5. \tag{12a}$$

Finally, Bernoulli numbers $B'_{2k} = 10^{-8k}B_{2k}$ were evaluated using the recursion relation

$$\binom{2k+1}{2k} B'_{2k} + 10^{-8}\binom{2k+1}{2k-2} B'_{2k-2} + \cdots + 10^{8-8k}\binom{2k+1}{2} B'_1 = (2k-1)/2 \cdot 10^{8k}. \tag{13}$$

From the fact that

$$\left| \frac{B_{2k}}{B_{2k-2}} \right| \approx \frac{2k(2k-1)}{4\pi^2} \tag{14}$$

it can be seen that the recursion (13) does not cause truncation errors to propagate. Furthermore, 1300 decimal places were used in all calculations.

When using (13) to calculate B'_{2k}, first all the positive terms were added together, then all the negative terms added together and finally the two were combined. This gave extra speed to the calculations. Care was also taken to avoid multiplying by zero. The evaluation of B'_{2k} becomes more difficult as k increases, because of the number of terms and the size of the binomial coefficients. Since the B_n alternate in sign, the actual error in the calculation of γ is less than $\frac{B'_{502}}{502} \approx +0.25 \times 10^{-1271}$, so the value obtained here should be correct to 1271 decimals. The fact that the final answer agreed with Adams' value and that numerous checks were made on all the arithmetical routines provides a good basis for guaranteeing the stated accuracy of the results. Dr. Wrench has independently verified the approximations to 1039 decimal places.

The present calculations were performed on a Burroughs 220 computer. The evaluation of S_{10000} required approximately one hour, and each of the logarithms required about six minutes. Evaluation of the 250 Bernoulli numbers was the most troublesome part of the calculations, and the total time for their calculation was approximately eight hours. A table of the Bernoulli numbers B' to 1270D has been sent to the Unpublished Mathematical Tables file of the journal, *Mathematics of Computation*.

4. Determination of Partial Quotients. To find best rational approximations to γ, we represent it as a continued fraction

$$\gamma = a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}. \tag{15}$$

Put $P_1 = Q_0 = 1$, $Q_1 = P_0 = 0$, and for $i \geqq 1$

$$P_{i+1} = a_i P_i + P_{i-1}\,. \tag{16}$$
$$Q_{i+1} = a_i Q_i + Q_{i-1}\,.$$

In matrix notation,

$$\begin{pmatrix} P_{i+1} & P_i \\ Q_{i+1} & Q_i \end{pmatrix} = \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_2 & 1 \\ 1 & 0 \end{pmatrix}\cdots\begin{pmatrix} a_i & 1 \\ 1 & 0 \end{pmatrix}. \tag{17}$$

Then $\lim P_i/Q_i = \gamma$. The fractions P_i/Q_i represent the best approximations to γ in the sense that

$$|\,Q_i\gamma - P_i\,| < |\,q\gamma - p\,|, \quad \text{if} \quad q < Q_i\,, \quad i \geqq 3. \tag{18}$$

We have then $a_i \neq 0$ $(i > 1)$, if and only if γ is irrational, and the sequence of partial quotients a_i will be periodic if and only if γ is quadratic, that is,

$$\gamma = r + \sqrt{s}, \quad r \text{ and } s \text{ rational.}$$

For proofs of these well-known results see Cassels [8].

The algorithm used to determine the partial quotients a_i, using limited decimal precision, is as follows:

Set $\gamma_1 = \gamma$, and

$$a_i = [\gamma_i], \quad \gamma_{i+1} = \{\gamma_i\}^{-1}, \quad i \geqq 1. \tag{19}$$

We have decimal numbers r_1 and s_1 such that

$$r_1 \leqq \gamma_1 \leqq s_1\,.$$

We successively find numbers r_i, s_i such that

$$r_i \leqq \gamma_i \leqq s_i\,. \tag{20}$$

If $[r_i] \neq [s_i]$, then the algorithm terminates. If $[r_i] = [s_i]$, then $[r_i] = a_i$ and

$$\{r_i\} \geqq \{\gamma_i\} \geqq \{s_i\}.$$

Hence

$$\{s_i\}^{-1} \leqq \gamma_{i+1} \leqq \{r_i\}^{-1}.$$

Choose decimal numbers r_{i+1} and s_{i+1} so that $r_{i+1} \leqq \{s_i\}^{-1}$ by truncation, $s_{i+1} \geqq \{r_i\}^{-1}$ by rounding up. Then the algorithm continues, until $[r_i] \neq [s_i]$.

The method used for calculating $\{s_i\}^{-1}$ when $\{s_i\}$ has several hundred decimal places was adapted from that of Pope and Stein [10]. Approximately six seconds was required to obtain each quotient. If t partial quotients are desired, the total time is proportional to t^3.

Table 1 gives the value of γ to 1271 decimal places. Table 2 gives the first 372 partial quotients of γ. Only 372 are given, although the value in Table 1 would have probably yielded over 1000 partial quotients. Table 3 gives for the reader's convenience the first few "best rational approximations" to γ. Here the ratio 228/395 gives a remarkably good value, correct to six decimal places.

From Table 2 one can compute

$$Q_{372} \approx 1.135 \times 10^{192}, \tag{21}$$

and we can conclude that if γ is rational its denominator must be larger than Q_{372}. Another consequence is that only about 385 decimal places of Table 1 were needed to obtain the 372 partial quotients. The referee has pointed out that Lehman [14] had already calculated the first 315 partial quotients for γ on the basis of Wrench's 328-place value [9]. These are in perfect agreement with the values obtained here.

The partial quotients of γ, as calculated in Table 2, appear to be "random" in some sense. Almost all real numbers have partial quotients satisfying

$$\lim \sqrt[n]{a_2 a_3 \cdots a_{n+1}} = K \tag{22}$$

where $K \approx 2.685$ is Khintchine's constant [13]. In this case,

$$\sqrt[371]{a_2 a_3 \cdots a_{372}} \approx 2.692,$$

a reasonable approximation to K.

TABLE 1

Value of Euler's Constant

.57721	56649	01532	86060	65120	90082	40243	10421	59335	93992
35988	05767	23488	48677	26777	66467	09369	47063	29174	67495
14631	44724	98070	82480	96050	40144	86542	83622	41739	97644
92353	62535	00333	74293	73377	37673	94279	25952	58247	09491
60087	35203	94816	56708	53233	15177	66115	28621	19950	15079
84793	74508	57057	40029	92135	47861	46694	02960	43254	21519
05877	55352	67331	39925	40129	67420	51375	41395	49111	68510
28079	84234	87758	72050	38431	09399	73613	72553	06088	93312
67600	17247	95378	36759	27135	15772	26102	73492	91394	07984
30103	41777	17780	88154	95706	61075	01016	19166	33401	52278
93586	79654	97252	03621	28792	26555	95366	96281	76388	79272
68013	24310	10476	50596	37039	47394	95763	89065	72967	92960
10090	15125	19595	09222	43501	40934	98712	28247	94974	71956
46976	31850	66761	29063	81105	18241	97444	86783	63808	61749
45516	98927	92301	87739	10729	45781	55431	60050	02182	84409
60537	72434	20328	54783	67015	17739	43987	00302	37033	95183
28690	00155	81939	88042	70741	15422	27819	71652	30110	73565
83396	73487	17650	49194	18123	00040	65469	31429	99297	77956
93031	00503	08630	34185	69803	23108	36916	40025	89297	08909
85486	82577	73642	88253	95492	58736	29596	13329	85747	39302
37343	88470	70370	28441	29201	66417	85024	87333	79080	56275
49984	34590	76164	31671	03146	71072	23700	21810	74504	44186
64759	13480	36690	25532	45862	54422	25345	18138	79124	34573
50136	12977	82278	28814	89459	09863	84600	62931	69471	88714
95875	25492	36649	35204	73243	64109	72682	76160	87759	50880
95126	20840	45444	77992	3(0)					

TABLE 2
Partial Quotients

	000	001	001	002	001	002	001	004	003	013	005	001	001	008
001	002	004	001	001	040	001	011	003	007	001	007	001	001	005
001	049	004	001	065	001	004	007	011	001	399	002	001	003	002
001	002	001	005	003	002	001	010	001	001	001	001	002	001	001
003	001	004	001	001	002	005	001	003	006	002	001	002	001	001
001	002	001	003	016	008	001	001	002	016	006	001	002	002	001
007	002	001	001	001	003	001	002	001	002	013	005	001	001	001
006	001	002	001	001	011	002	005	006	001	001	001	006	001	002
002	001	005	006	002	001	001	007	013	004	001	002	004	001	004
001	001	023	001	009	005	002	001	001	001	008	003	002	004	002
033	005	001	002	001	003	002	004	002	001	005	012	001	017	006
002	032	005	003	001	006	001	003	001	002	001	018	001	002	017
001	006	001	021	001	006	001	071	018	001	006	058	002	001	013
055	001	103	001	014	001	005	008	001	002	010	002	001	001	003
003	002	001	182	001	004	003	002	004	001	002	001	001	001	006
001	001	001	006	001	003	002	069	002	001	006	002	002	012	001
001	001	008	001	002	003	002	001	052	001	025	004	002	018	001
040	001	018	001	002	014	001	002	002	010	001	001	002	006	071
007	001	010	002	001	001	001	002	001	003	002	004	001	006	003
001	001	029	001	029	001	001	003	004	007	001	001	010	002	002
030	001	021	003	012	001	039	008	007	001	002	001	002	002	001
001	002	003	001	013	001	002	003	001	001	001	001	008	007	001
001	001	004	002	005	012	001	015	005	001	007	001	005	001	001
001	006	005	001	041	001	005	001	009	013	001	001	005	021	025
008	005	001	014	001	001	001	006	003	001	100	001	265		

TABLE 3
Best Rational Approximations

1/2	.50
3/5	.60
4/7	.571
11/19	.579
15/26	.5769
71/123	.57724
228/395	.5772152
3035/5258	.57721567
15403/26685	.5772156642
18438/31943	.5772156654
33841/58628	.57721566487

The author wishes to acknowledge his gratitude to the Burroughs Corporation and to the Case Institute of Technology for the use of their Burroughs 220 computers.

California Institute of Technology
Pasadena, California

1. L. EULER, "De progressionibus harmonicis observationes," *Euleri Opera Omnia Ser. 1*, v. 14, Teubner, Leipzig and Berlin, 1925, p. 93–100.
2. L. EULER, "De summis serierum numeros Bernoullianos involventium," *Euleri Opera Omnia Ser. 1*, v. 15, Teubner, Leipzig and Berlin, 1927, p. 91–130. See especially p. 115. The calculation is given in detail on p. 569–583.

3. J. W. L. GLAISHER, "On the calculation of Euler's constant," *Proc. Roy. Soc. London*, v. 19, 1870, p. 514–524.
4. W. SHANKS, "On the numerical value of Euler's constant," *Proc. Roy. Soc. London*, v. 15, 1867, p. 429–432; v. 20, 1871, p. 29–34.
5. J. W. L. GLAISHER, "History of Euler's constant," *Messenger of Mathematics*, v. 1, 1872, p. 25–30.
6. J. C. ADAMS, "On the value of Euler's constant," *Proc. Roy. Soc. London*, v. 27, 1878, p. 88–94. See also v. 42, 1887, p. 22–25.
7. K. KNOPP, *Theory and Application of Infinite Series*, Blackie and Son, London, 1951, p. 257.
8. J. W. S. CASSELS, *An Introduction to Diophantine Approximation*, Cambridge University Press, 1957, p. 1–11.
9. J. W. WRENCH, JR., "A new calculation of Euler's constant," *MTAC*, v. 6, 1952, p. 255.
10. D. A. POPE & M. L. STEIN, "Multiple precision arithmetic," *Comm. ACM*, v. 3, 1960, p. 652–654.
11. D. SHANKS & J. W. WRENCH, JR., "Calculation of π to 100,000 decimals," *Math. Comp.*, v. 16, 1962, p. 76–99.
12. D. J. WHEELER, *The Calculation of 60,000 Digits of e by the Illiac*, Digital Computer Laboratory Internal Report No. 43, University of Illinois, Urbana, 1953.
13. J. W. WRENCH, JR., "Further evaluation of Khintchine's constant," *Math. Comp.*, v. 14, 1960, p. 370–371.
14. R. S. LEHMAN, *A Study of Regular Continued Fractions*, Ballistic Research Lab. Report 1066, Aberdeen Proving Ground, Maryland, February 1959.

A Calculation of the Number of Lattice Points in the Circle and Sphere

W. Fraser and C. C. Gotlieb

1. Introduction. Let $A_k(x)$ be the number of points $(y_1, y_2, \cdots y_k)$ satisfying

$$y_1^2 + y_2^2 + \cdots y_k^2 \leqq x \tag{1}$$

where $y_1, y_2 \cdots y_k$ are integers (positive, negative or zero). Thus, $A_k(x)$ is the number of lattice points in a k-dimensional hypersphere of radius $x^{1/2}$. This paper describes the calculation of a table of $A_2(x)$ and $A_3(x)$ on an IBM 650 computer.

As a first approximation these are, respectively, the area and volume of the circle and sphere, but the question is how good these approximations are. In general, we are interested in

$$P_k(x) = A_k(x) - V_k(x) \tag{2}$$

where the volume of a sphere of radius $x^{1/2}$ in k-dimensional hyperspace is

$$V(x) = \frac{\Pi^{k/2} x^{k/2}}{\Gamma(k/2 + 1)}. \tag{3}$$

$P_2(x)$ has been investigated by many celebrated mathematicians and Wilton [1] gives an account of the early work. More recently there have been theoretical investigations of $P_k(x)$ for higher dimensions, particularly by Walfisz [2], whose notation is being followed here.

We write $P_2(x) = O(x^c)$ to mean, in the usual sense, that there exists K such that

$$\begin{aligned} |P_2(x)|/x^c &\leqq K \quad \text{as } x \to \infty, \text{ and} \\ P_2(x) &= o(x^c) \quad \text{to mean that} \\ P_2(x)/x^c &\to 0 \quad \text{as } x \to \infty. \end{aligned}$$

Further, after Littlewood, we write $P_2(x) = \Omega(x^c)$ to mean that there exists $K > 0$, and a sequence of values of x tending to infinity, for which

$$|P_2(x)|/x^c \geqq K,$$

that is, the negation of $P_2(x) = o(x^c)$. Gauss observed that

$$P_2(x) = O(x^{1/2})$$

Hua [3] has shown that

$$P_2(x) = O(x^{13/40}),$$

and Van der Corput [4] that

$$P_2(x) = o(x^{1/3})$$

Received August 7, 1961.

TABLE 1
Contributions to $A_3(x)$—First Method

Name	Solutions for:	Conditions	Symmetry Factors		Multiplicity
			Permutation	Reflection	
T	$y_1^2 + y_2^2 + y_3^2 \leqq x$	$y_1 > y_2 > y_3 > 0$	6	8	48
U	$y_1^2 + y_2^2 \leqq x$	$y_1 > y_2 > 0$	6	4	24
V	$2y_1^2 + y_2^2 \leqq x$	$y_1 > y_2 > 0$	3	8	24
W	$y_1^2 + 2y_2^2 \leqq x$	$y_1 > y_2 > 0$	3	8	24
S	$y_1^2 \leqq x$	$y_1 > 0$	3	2	6
Y	$2y_1^2 \leqq x$	$y_1 > 0$	3	4	12
Z	$3y_1^2 \leqq x$	$y_1 > 0$	1	8	8

Hardy has shown that

$$P_2(x) = \Omega(x^{1/4}),$$

so that we may write

$$P_2(x) = O(x^c)$$

where $c \geqq \frac{1}{4}$ and the best value is known to be less than 13/40. It is considered probable that c is arbitrarily close to $\frac{1}{4}$. Besides furnishing numerical values of $A_2(x)$ and $A_3(x)$, one purpose of this tabulation is to examine the consistency of the observed numbers with this conjecture.

Although analytic number theory yields considerable information on the behavior of $P_k(x)$ for $k \geqq 4$, there seems to be little reported on $P_3(x)$. In general

$$P_k(x) = O(x^{(k-1)/2}) \quad \text{and} \quad P_k(x) = \Omega(x^{k/2-1}) \tag{4}$$

so that

$$P_3(x) = O(x^c)$$

where it is known that $c \geqq \frac{1}{2}$ and the best value is equal to or less than 1. These limits, of course, are not very sharp; for $P_2(x)$ we get from (4) only that $\frac{1}{2} \geqq c \geqq 0$. Thus, another purpose of the tabulation is to obtain some information about the behavior of $P_3(x)$.

2. Computing Formulas. There are many summation formulas which can be derived for $A_3(x)$. Essentially they are all modified enumerations where advantage is taken of the symmetries present.

Table 1 shows a decomposition into the terms which contribute to $A_3(x)$. Allowing for the solution at the origin, we then get

$$A_3(x) = 48T + 24(U + V + W) + 6S + 12Y + 8Z + 1^* \tag{5}$$

* As noted by Legendre this same result is reached by noting that $A_3(x)$ is given by the number of terms having coefficients $\leqq x$ in the expansion of $(1 + 2\sum_i y^{i^2})(1 + 2\sum_j y^{j^2})(1 + 2\sum_k y^{k^2})$.

We proceed to find expressions for each of these terms. Let $[\sqrt{N}]$ be the largest integer equal to or less than the square root of N, i.e.,

$$[\sqrt{N}] \leqq \sqrt{N} < [\sqrt{N}] + 1.$$

Then

$$S = [\sqrt{x}]; \qquad Y = [\sqrt{x/2}]; \qquad Z = [\sqrt{x/3}].$$

The other terms in (5) are evaluated by summation formulas. Consider the number of solutions of

$$y_1^2 + y_2^2 \leqq x \quad \text{with} \quad y_1 \neq 0, \quad y_2 \neq 0. \tag{6}$$

For each y_1, the number of permitted values of y_2 is $[\sqrt{x - y_1^2}]$, so that the number of solutions of (6) is given by $\sum_{y_1=1}^{[\sqrt{x}]} [\sqrt{x - y_1^2}]$. But this is equal to

the number of solutions with $y_1 > y_2 > 0$

+ the number of solutions with $y_2 > y_1 > 0$

+ the number of solutions with $y_1 = y_2$.

That is,

$$2U + Y = \sum_{y_1=1}^{[\sqrt{x}]} [\sqrt{x - y_1^2}].$$

A similar argument shows that

$$V + W = \sum_{y_1=1}^{[\sqrt{x/2}]} [\sqrt{x - 2y_1^2}] - [\sqrt{x/3}];$$

again, we have the number of solutions of

$$y_1^2 + y_2^2 + y_3^2 \leqq x \quad \text{for} \quad y_1 > y_2 > 0 = 3T + V + w$$

$$= \sum_{y_1=2}^{[\sqrt{x-1}]} \sum_{y_2=1}^{\min} [\sqrt{x - y_1^2 - y_2^2}]$$

where min is the lesser of $(y_1 - 1)$ and $[\sqrt{x - y_1^2}]$. Thus we finally get

$$A_3(x) = 16 \sum_{y_1=1}^{[\sqrt{x-1}]} \sum_{y_2=1}^{\min} [\sqrt{x - y_1^2 - y_2^2}] + 8 \sum_{y_1=1}^{[\sqrt{x/2}]} [\sqrt{x - 2y_1^2}] + 12 \sum_{y_1=1}^{[\sqrt{x}]} [\sqrt{x - y_1^2}] + 6[\sqrt{x}] + 1. \tag{7}$$

Further, it is easily seen that

$$A_2(x) = 4 \sum_{y_1=1}^{[\sqrt{x}]} [\sqrt{x - y_1^2}] + 4[\sqrt{x}] + 1. \tag{8}$$

These equations, (7) and (8), were the two actually used the first time the table was computed. If it had been practical to compute in successive values of x it probably would have been best to use a difference formula to find the contributions from successive spherical shells. Actually it was decided to compute in equal intervals

of $x^{1/2}$ for a while, and then to skip to large arguments so that the asymptotic values could be examined. The IBM 650 computer used for the calculation was equipped with core storage and index registers. Although the full program required only 168 instructions and the inner loop was completely in the 60-word core storage, it took about 10 hours to compute $A_3(10^6)$, the largest value for which equation (7) was used.

A second and independent program was written for computing $A_2(x)$ and $A_3(x)$. This program was based on the equations:

$$A_3(x) = 1 + 6[\sqrt{x}] + 12\left[\sqrt{\frac{x}{2}}\right] + 8\left[\sqrt{\frac{x}{3}}\right] + 24\sum_{y_1=1}^{L_1}([\sqrt{x-y_1^2}] - y_1)$$

$$+ 24\sum_{y_1=1}^{L_2}([\sqrt{x-2y_1^2}] - y_1) + 24\sum_{y_1=1}^{L_3}\left(\left[\sqrt{\frac{x-y_1^2}{2}}\right] - y_1\right) \tag{9}$$

$$+ 48\sum_{y_1=1}^{L_4}\sum_{y_2=y_1+1}^{L_5}([\sqrt{x-y_1^2-y_2^2}] - y_2),$$

where

L_1 is the largest integer, y_1 satisfying $y_1^2 + (y_1+1)^2 \leqq x$

L_2 is the largest integer, y_1 satisfying $2y_1^2 + (y_1+1)^2 \leqq x$

L_3 is the largest integer, y_1 satisfying $y_1^2 + 2(y_1+1)^2 \leqq x$

L_4 is the largest integer, y_1 satisfying $y_1^2 + (y_1+1)^2 + (y_1+2)^2 \leqq x$

L_5 is the largest integer, $y_2(y_1)$ satisfying $y_1^2 + y_2^2 + (y_1+1)^2 \leqq x$;

$$A_2(x) = 1 + 4[\sqrt{x}] + 4[\sqrt{x/2}] + 8\sum_{y_1=1}^{L_0}([\sqrt{x-y_1^2}] - y_1) \tag{10}$$

where L_0 is the largest integer, y_1 satisfying $y_1^2 + (y_1+1)^2 \leqq x$. These equations can be obtained by counting the lattice points according to the scheme illustrated in Figure 1 and Table 2.

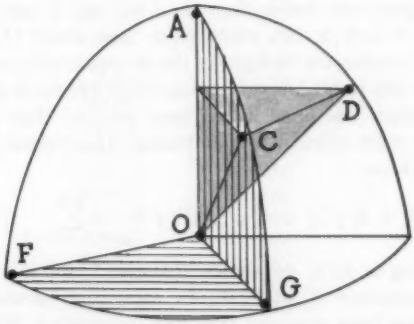


FIG. 1. Decomposition of $A_2(x)$—second method.

TABLE 2
Contributions to $A_2(x)$—Second Method

Source	Multiplicity	Number of Points
Point O	1	1
Line OF	6	$[\sqrt{x}]$
Line OG	12	$[\sqrt{x/2}]$
Line OC	8	$[\sqrt{x/3}]$
Plane OFG	24	$\sum([\sqrt{x - y_1^2}] - y_1)$
Plane OCA	24	$\sum([\sqrt{x - 2y_1^2}] - y_1)$
Plane OCG	24	$\sum\left(\left[\sqrt{\frac{x - y_1^2}{2}}\right] - y_1\right)$
Volume OACD	48	$\sum\sum([\sqrt{x - y_1^2 - y_2^2}] - y_2)$

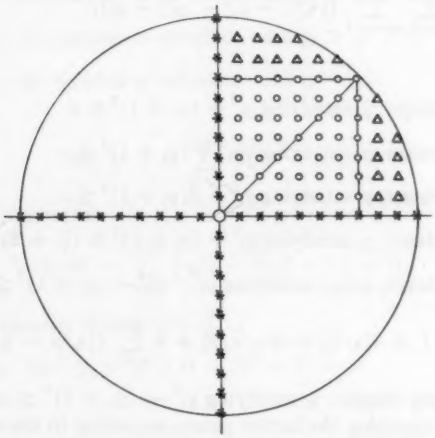


FIG. 2. Decomposition of $A_2(x)$—Gauss method.

This second program was faster than the first, and it was possible to extend the calculation to $x = 3.24 \times 10^6$, which value took about $1\frac{1}{2}$ hours to compute. For this value, $A_3(x)$ exceeds the 10 digits of the storage positions, but the individual sums contributing to the result are still within single-precision range.

Since the calculation was done it has been realized that there are formulas which might be even more efficient for computing. These result from the following equation, noted by Gauss:

$$A_2(x) = 1 + 4[\sqrt{x}] + 4[\sqrt{x/2}]^2 + 8 \sum_{y_1 = [\sqrt{x/2}]+1}^{[\sqrt{x}]} [\sqrt{x - y_1^2}]. \tag{11}$$

The terms contributing to $A_2(x)$ are shown in Figure 2, and it is seen that equation (11) is superior to equations (8) and (10) for computing because the contribution given by $4[\sqrt{x/2}]^2$ has been removed from the summation. When equation (11)

TABLE 3
Computed Results

$X^{1/2}$	$A_3(X)$	$\lvert P_3(X) \rvert$	$A_2(X)$	$\lvert P_2(X) \rvert$
1	7	3	5	2
2	33	1	13	0
3	123	10	29	1
4	257	11	49	1
5	515	9	81	2
6	925	20	113	0
7	1419	18	149	5
8	2109	36	197	4
9	3071	17	253	1
10	4169	20	317	3
11	5575		377	3
12	7153	85	441	11
13	9171	32	529	2
14	11513	19	613	3
15	14147	10	709	2
16	17077	80	797	7
17	20479	101	901	7
18	24405	24	1009	9
19	28671	60	1129	5
20	33401	109	1257	0
21	38911	119	1373	12
22	44473	129	1517	4
23	50883	82	1653	9
24	57777	129	1793	17
25	65267	183	1961	2
26	73525	97	2121	3
27	82519	71	2289	1
28	91965	13	2453	10
29	101943	217	2629	13
30	113081	16	2821	6
31	124487	301	3001	18
32	137065	193	3209	8
33	150555	22	3409	12
34	164517	119	3625	7
35	179579	15	3853	5
36	195269	163	4053	19
37	212095	80	4293	8
38	229549	298	4513	23
39	248439	36	4777	1
40	267761	322	5025	2
41	288359	337	5261	20
42	310177	162	5525	17
43	332779	259	5789	20
44	356637	181	6077	5
45	381915	211	6361	1
46	407597	123	6625	23
47	434551	342	6921	19
48	462781	466	7213	25
49	492567	240	7525	18

TABLE 3—*Continued*

$X^{1/2}$	$A_3(X)$	$\lvert P_3(X)\rvert$	$A_2(X)$	$\lvert P_2(X)\rvert$
50	523305	294	7845	9
55	696507	403	9477	26
60	904089	690	11289	21
65	1149651	696	13273	0
70	1436385	370	15373	21
75	1767063	83	17665	6
80	2143641	1020	20081	25
85	2571711	730	22701	3
90	3053617	11	25445	2
95	3590863	501	28345	8
100	4187857	933	31417	1
105	4849327	279	34621	15
110	5574721	559	37981	32
115	6370351	275	41545	3
120	7236577	1652	45225	14
125	8180887	344	49077	10
130	9201625	1147	53077	16
135	10305407	588	57209	47
140	11492081	1959	61529	46
145	12768503	1548	66045	7
150	14137637	470	70681	5
155	15598031	500	75465	12
160	17155325	1960	80381	44
165	18817007	438	85501	29
170	20578325	1201	90785	7
175	22448927	371	96209	2
180	24427317	1707	101765	23
185	26520663	1186	107501	20
190	28729653	1259	113369	42
195	31058271	1085	119433	26
200	33507885	2437	125629	35
300	113094545	2791	282697	46
400	268077737	4836	502625	30
500	523592077	6699	785349	49
600	904769241	9443	1130913	60
700	1436743985	11055	1539297	83
800	2144654669	5916	2010573	46
900	3053616505	11554	2544569	121
1000	4188781437	8768	3141549	44
1200	7238202017	27457	4523793	100
1400	11494026189	14133	6157477	145
1600	17157266213	18466	8042349	128
1800	24428980617	43857	10178545	215

is integrated plane-by-plane there results

$$A_3(x) = 1 + 6[\sqrt{x}] + 12[\sqrt{x/2}] + 24 \sum_{y_1=[\sqrt{x/2}]+1}^{[\sqrt{x}]} [\sqrt{x - y_1^2}]$$

$$+ 8 \sum_{y_1=1}^{[\sqrt{x-1}]} \left[\sqrt{\frac{x - y_1^2}{2}}\right] + 16 \sum_{y_1=1}^{[\sqrt{x-1}]} \sum_{y_2=\left[\sqrt{\frac{x-y_1^2}{2}}+1\right]}^{[\sqrt{x-y_1^2}]} [\sqrt{x - y_1^2 - y_2^2}]. \tag{12}$$

Still other summation formulas can be derived and it is likely that the most efficient one will depend on the computer being used. A method of decomposition for finding $A_k(x)$ which is best for $k = 2$ is not necessarily best for higher values of k.

3. Results. A partial table of computed results is shown in Table 3. Shown there are $x^{1/2}$, $A_3(x)$, $|P_3(x)|$, $A_2(x)$ and $|P_2(x)|$ for the following range:

$$x^{1/2} = 1(1)\ 50(5)\ 200(100)\ 1000(200)\ 1800.$$

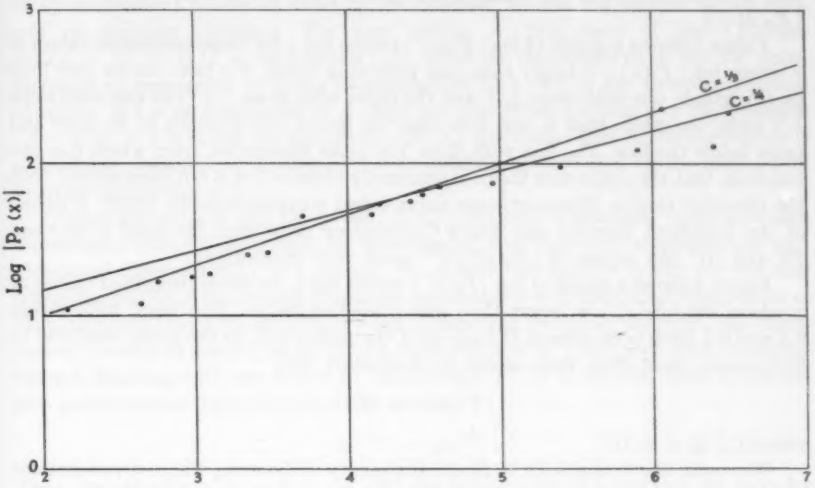


FIG. 3. Log X.

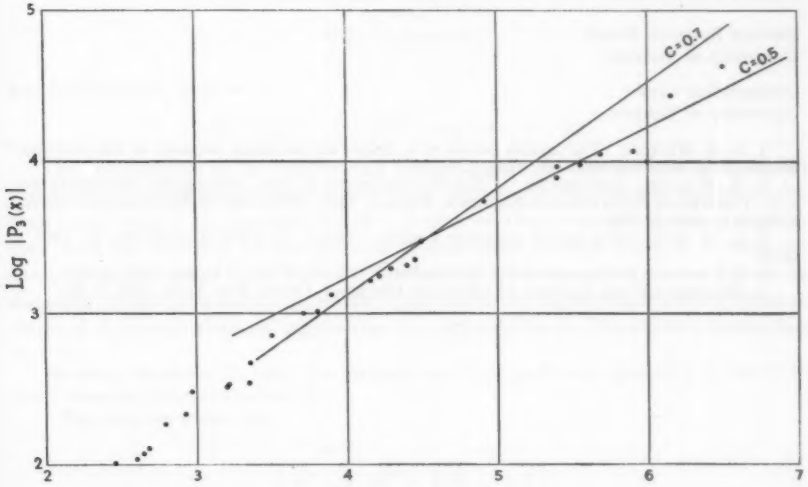


FIG. 4. Log X.

All of these entries except those for $x^{1/2} = 1200, 1400, 1600$ and 1800 were computed by the two independent programs. The results agree with those given in the short table of reference [5]. Calculations were actually made for about 250 arguments in the range $x^{1/2} = 1$ to 2000.

Because it is known that $P_2(x) = \Omega(x^{1/4})$ the points of greatest interest in investigating the asymptotic behavior of $P_2(x)$ are those local maxima, M_i of $| P_2(x) |$, satisfying $| P_2(M_i) | > | P_2(x) |$ for all $x < M_i$. From the tabulated results it is only possible to obtain an estimate of these M_i and the corresponding values of $| P_2(M_i) |$.

Figure 3 shows a graph of $\log | P_2(x) |$ versus $\log x$ for those computed values of x where $\log | P_2(x) |$ is larger than any preceding value. We have drawn two lines on this graph, one with slope 1/3, and the other with slope 1/4. The line with slope 1/3 looks too steep, that is, one feels that the points will continue to lie more and more below the line. The line with slope 1/4 looks reasonable, from which one can conclude that the conjecture that c is arbitrarily close to 1/4 is not inconsistent with the observed results. However some unpublished computations by Harry Mitchell of the Lockheed Missiles and Space Corporation show that for some x between 10^8 and 10^{10} the values of $| P_2(x) |/x^{1/4}$ grow very remarkably.

Figure 4 shows a graph of $\log | P_3(x) |$ versus $\log x$, for those computed values of x where $\log | P_3(x) |$ is larger than any preceding value. Two lines, with slopes 0.5 and 0.7 have been drawn. It looks as if the points will, in the main, continue to lie between these lines, from which we conjecture that

$$| P_3(x) | = O(x^c)$$

where $0.5 \leqq c \leqq 0.7$.

We would like to thank D. B. Scott, University of Toronto, whose investigations into close packing of spheres first suggested this calculation, and G. Forsythe, Stanford University, with whom we have had helpful discussions.

Defence Research Board
University of Toronto

Computation Center
University of Toronto

1. J. R. Wilton, "The lattice points of a cirlce: an historical account of the problem," *Messenger of Mathematics*, v. 58, 1929, p. 67–80.
2. A. Walfisz, *Gitterpunkte in Mehrdimensionalen Kugeln*, Monografie Matematyczne, v. 33, Panstwowe Wydawnictwo Naukowe, Warsaw, 1957. (Reviewed in *Mathematical Reviews*, v. 20, n. 6, 1959, p. 3826.)
3. L. K. Hua, "The lattice-points in a circle," *Quart. J. Math., Oxford Ser.* 13, 1942, p. 18–29.
4. E. Landau, *Vorlesungen ueber Zahlentheorie*, Von S. Hirzel, Leipzig, 1927, p. 264.
5. Hilbert & Cohn-Vossen, *Anschauliche Geometrie*, Dover, New York, 1944. p. 29.

Sign Wave Analysis in Matrix Eigenvalue Problems

By K. M. Brown and P. Henrici

1. Introduction. In this paper the word "matrix" denotes a real (but not necessarily symmetric) matrix of order N; by "vector" we mean a column vector with real or complex elements. For any matrix A, the roots of the equation $\det(A - \lambda I) = 0$ (I = unit matrix) are called the *eigenvalues* of A. By the *multiplicity* of an eigenvalue we mean its multiplicity as a root of the above polynomial equation. If λ is an eigenvalue of A, then any nontrivial solution x of the equation $(A - \lambda I)x = 0$ is called an eigenvector (of A) associated with λ. An eigenvalue is a *dominant eigenvalue of the matrix* A if its modulus is exceeded by the modulus of no other eigenvalue of A.

The power method ([1]; [3], p. 296; [5]; [7]; [9]; [10]) is generally recognized as a numerically efficient algorithm for determining the dominant eigenvalue(s) and associated eigenvector(s) of a matrix. We review the method briefly for the case where the matrix A has a single dominant eigenvalue λ with associated eigenvector u. (It is assumed that λ has multiplicity one.) Denoting by the superscript ^{T} the transpose of a vector or matrix, we let v be an eigenvector of A^T associated with the eigenvalue λ. Starting with any vector $x^{(0)}$ satisfying $v^T x^{(0)} \neq 0$, we now form by successive matrix-vector multiplications the vectors

$$x^{(n+1)} = Ax^{(n)}, \qquad n = 0, 1, 2, \cdots.$$

Then, denoting by a_ν ($\nu = 1, 2, \cdots, N$) the components of a vector a, we have for every ν such that $u_\nu \neq 0$

$$\lim_{n\to\infty} \frac{x_\nu^{(n+1)}}{x_\nu^{(n)}} = \lambda,$$

and furthermore, as $n \to \infty$,

$$x_1^{(n)} : x_2^{(n)} : \cdots : x_N^{(n)} \to u_1 : u_2 : \cdots u_N .^*$$

The convergence of the process can be sped up by devices such as shift of the origin [10], fractional iteration [8], and the δ^2-process [1]. Statements similar to the above still hold if the multiplicity of λ is greater than one, but the convergence may then be slow due to the presence of nonlinear divisors. Once λ, u, and the associated eigenvector of A^T have been determined, one can, by a process known as deflation, construct a matrix A_1 whose eigenvalues and eigenvectors are the same as those of A, except that the eigenvalue λ is replaced by 0. The above process can

Received December 17, 1961. The preparation of this paper was sponsored by the U. S. Army Research Office at Durham, N. C.

* This notation means that

$$\frac{x_\nu^{(n)}}{\left[\sum_{\mu=1}^{N} |x_\mu^{(n)}|^2\right]^{1/2}} \to \frac{u_\nu}{\left[\sum_{\mu=1}^{N} |u_\mu|^2\right]^{1/2}} \qquad (n \to \infty, \nu = 1, 2, \cdots, N).$$

then be repeated; if the matrices $A_1, A_2, \cdots$ all have single dominant eigenvalues, the above method yields successively all eigenvalues and eigenvectors of the matrix A.

2. Conjugate Complex Dominant Eigenvalues. In the present note we wish to deal with the case where the matrix A has exactly two dominant eigenvalues, both simple, represented by the pair of conjugate complex numbers $\lambda = \rho e^{i\varphi}$ and $\bar{\lambda} = \rho e^{-i\varphi}$, where $\rho > 0, 0 < \varphi < \pi$. The eigenvectors associated with λ and with $\bar{\lambda}$ may then be assumed to be conjugate complex vectors also. We shall denote them by u and $\bar{u}$, where the components u_ν of u are given by

$$u_\nu = r_\nu e^{i\varphi_\nu}, \qquad \nu = 1, 2, \cdots, N.$$

In view of the fact that the eigenvectors are determined only up to a non-zero factor (which in the present case may even be complex), it should be noted that the r_ν are determined only up to a positive factor, and the φ_ν only up to a common additive constant modulo 2π. Indicating by the superscript ^{H} the conjugate transpose of a complex vector or matrix, we denote by v the eigenvector of A^H belonging to $\bar{\lambda}$, normalized such that $v^H u = 1$.

One of the methods for determining conjugate complex eigenvalues and corresponding eigenvectors from the sequence $\{x^{(n)}\}$ that have been proposed ([3], p. 296; [9]) is known to be numerically unstable for small values of φ [9]. In Section 3 below we propose an alternate method that appears to be uniformly accurate for all values of φ. In addition, the method yields very good approximations for both φ and the φ_ν almost without computation, by mere inspection of the signs of the sequences of the components of the (real) vectors $x^{(n)}$.

3. Sign Waves. It is known ([2], p. 285) that the presence of a pair of conjugate complex dominant eigenvalues is indicated by the occurrence of sign changes in the sequences $\{x_\nu^{(n)}\}$. For a certain matrix of order 6, the signs of the $x_\nu^{(n)}$ were distributed as follows:

ν	1	2	3	4	5	6
$n = 0$	+	+	+	+	+	+
1	+	+	+	−	+	+
2	−	−	−	−	−	+
3	−	−	−	+	−	−
4	+	+	−	+	+	−
5	+	+	+	−	+	+
6	−	−	+	−	−	+
7	−	−	−	+	−	−
8	−	−	−	+	−	−
9	+	+	+	−	+	+
10	+	+	+	−	+	+
⋮						

Our method consists in exploring systematically the period and relative location of these *sign waves*. We ascribe a sign to *all* elements of the N sequences

$\{x_\nu^{(n)}\}$ ($\nu = 1, \cdots, N$). A zero element is assigned the sign of the first nonzero element following it. (If there is no such element, the sign is irrelevant for the following theory.) For $k = 1, 2, \cdots$ we denote by $n_\nu^{(k)}$ the index of the element in the sequence $\{x_\nu^{(n)}\}$ at which the sign changes from minus to plus for the kth time, i.e., at which

$$\operatorname{sign} x_\nu^{(n-1)} = -1, \qquad \operatorname{sign} x_\nu^{(n)} = +1.$$

(The indices $n_\nu^{(k)}$ mark the beginnings of the kth sign wave in the sequence $\{x_\nu^{(n)}\}$. In the example given above, $n_1^{(1)} = 4$, $n_1^{(2)} = 9$, $n_5^{(2)} = 9$.) For $\nu, \mu = 1, 2, \cdots, N$ and $k = 1, 2, \cdots$ we put

$$P_\nu^{(k)} = n_\nu^{(k+1)} - n_\nu^{(k)},$$

$$\delta_{\nu\mu}^{(k)} = n_\mu^{(k)} - n_\mu^{(k)}.$$

($P_\nu^{(k)}$ indicates the length of the kth sign wave in the νth component, and $\delta_{\nu\mu}^{(k)}$ represents the phase difference between the kth sign waves in the νth and μth components. In the above example, $P_3^{(1)} = 4$, $\delta_{42}^{(2)} = -2$.) We finally require the quantities

$$\Delta_\nu^{(n)} = [x_\nu^{(n)}]^2 - x_\nu^{(n+1)} x_\nu^{(n-1)}.$$

With these definitions, we can state the following result:

THEOREM. *Let the matrix A satisfy the conditions stated at the beginning of Section 2, and let the vector $x^{(0)}$ be such that $v^H x^{(0)} \neq 0$. Then the vectors $x^{(n)}$ defined by (1) satisfy*

(i) $$\Delta_1^{(n)} : \Delta_2^{(n)} : \cdots : \Delta_N^{(n)} \to r_1^2 : r_2^2 : \cdots : r_N^2.$$

For every ν such that $r_\nu \neq 0$, the following two statements hold:

(ii) $$\rho^2 = \lim_{n\to\infty} \frac{\Delta_\nu^{(n+1)}}{\Delta_\nu^{(n)}};$$

(iii) $$\lim_{k\to\infty} \frac{P_\nu^{(1)} + P_\nu^{(2)} + \cdots + P_\nu^{(k)}}{k} = P$$

exists, and

$$\varphi = \frac{2\pi}{P}.$$

For all ν and μ such that $r_\nu \neq 0$, $r_\mu \neq 0$, the following two statements hold:

(iv) *If $\varphi/2\pi$ is irrational, then*

$$\lim_{k\to\infty} \frac{\delta_{\nu\mu}^{(1)} + \delta_{\nu\mu}^{(2)} + \cdots + \delta_{\nu\mu}^{(k)}}{kP} \equiv \frac{\varphi_\mu - \varphi_\nu}{2\pi} \pmod 1$$

(v) *If $\varphi/2\pi = p/q$ is rational ($(p, q) = 1$) it can only be asserted that, for some integer l, both the limit superior and the limit inferior of*

$$\frac{\delta_{\nu\mu}^{(1)} + \delta_{\nu\mu}^{(2)} + \cdots + \delta_{\nu\mu}^{(k)}}{kP}$$

as $k \to \infty$ differ by at most $1/q$ from

$$\frac{\varphi_\mu - \varphi_\nu}{2\pi} + l.$$

4. Proof of the Theorem. Under the hypotheses of the theorem the matrix A can be represented in the form

$$A = \lambda u w^T + \bar{\lambda}\bar{u}v^H + A_1$$

where A_1 is a matrix whose eigenvalues lie inside a circle of radius $q\rho$, $0 < q < 1$. If $v^T x^{(0)} = ae^{i\alpha}$, where $a > 0$, we have

$$x^{(n)} = a(\lambda^n u e^{i\alpha} + \bar{\lambda}^n \bar{u} e^{-i\alpha} + w^{(n)})$$

where the components of $w^{(n)}$ are bounded by $Cq^n\rho^n$ with a suitable constant C. Hence

$$x_\nu^{(n)} = 2a\rho^n\{r_\nu \cos(n\varphi + \varphi_\nu + \alpha) + \epsilon_\nu^{(n)}\}, \tag{1}$$

where

$$|\epsilon_\nu^{(n)}| \leqq Cq^n. \tag{2}$$

A simple calculation now yields

$$\Delta_\nu^{(n)} = \frac{4a^2\rho^{2n}}{(\sin\varphi)^2}(r_\nu^2 + \eta_\nu^{(n)}),$$

where

$$|\eta_\nu^{(n)}| \leqq 2Cq^n \sin^2\varphi(2r_\nu + Cq^n)$$

and hence $\eta_\nu^{(n)} \to 0$ as $n \to \infty$. The relations (i) and, if $r_\nu \neq 0$, (ii) now follow immediately.

For the proof of the remaining statements of the theorem a modified version of (1) is required. Assume the integer n' is such that, for all ν satisfying $r_\nu \neq 0$, $|\epsilon_\nu^{(n)}| < r_\nu$ for $n \geqq n'$. Setting temporarily $\beta_\nu^{(n)} = n\varphi + \varphi_\nu + \alpha$, we then have for $n \geqq n'$

$$\begin{aligned}(2a\rho^n)^{-1}x_\nu^{(n)} &= \operatorname{Re}\{r_\nu e^{i\beta_\nu^{(n)}} + \epsilon_\nu^{(n)}\} \\ &= \operatorname{Re}\{e^{i\beta_\nu^{(n)}}[r_\nu + e^{-i\beta_\nu^{(n)}}\epsilon_\nu^{(n)}]\} \\ &= |r_\nu + e^{-i\beta_\nu^{(n)}}\epsilon_\nu^{(n)}| \operatorname{Re}\{e^{i(\beta_\nu^{(n)} - \theta_\nu^{(n)})}\},\end{aligned}$$

where

$$\tan\theta_\nu^{(n)} = \frac{\epsilon_\nu^{(n)} \sin\beta_\nu^{(n)}}{r_\nu + \epsilon_\nu^{(n)}\cos\beta_\nu^{(n)}}, \qquad |\theta_\nu^{(n)}| < \frac{\pi}{2}. \tag{3}$$

Hence

$$x_\nu^{(n)} = A_\nu^{(n)} \sin\phi_\nu^{(n)} \tag{4}$$

where

$$A_\nu^{(n)} = 2a\rho^n |r_\nu + e^{-i\beta_\nu^{(n)}}\epsilon_\nu^{(n)}| > 0,$$

$$\phi_\nu^{(n)} = n\varphi + \varphi_\nu + \alpha + \frac{\pi}{2} - \theta_\nu^{(n)}.$$

Formula (4) serves to determine the sign of $x_\nu^{(n)}$ as a function of n.

We shall require an explicit formula for $n_\nu^{(k)}$ valid for large k. It follows from (2) and (3) that $\theta_\nu^{(n)} \to 0$ for $n \to \infty$. Let n'' be such that

$$|\theta_\nu^{(n)}| \leqq \tfrac{1}{2}\,\mathrm{Min}\,(\varphi, \pi - \varphi), \qquad n > n''$$

for all ν satisfying $r_\nu \neq 0$. We then have

$$0 < \phi_\nu^{(n+1)} - \phi_\nu^{(n)} < \pi, \qquad n > n'', \tag{5}$$

i.e., the sequence $\{\phi_\nu^{(n)}\}$ is monotonically increasing and assumes a value in every (open) interval of length π. Let k_ν denote the smallest integer such that

$$n_\nu^{(k_\nu)} > n''.$$

By (5) and by the definition of $n_\nu^{(k)}$, there exists an integer m_ν such that

$$\phi^{(n(k_\nu))} = n_\nu^{(k_\nu)}\varphi + \varphi_\nu + \alpha + \frac{\pi}{2} - \theta_\nu^{(n)} \geqq 2m_\nu\pi,$$

$$\phi_\nu^{(n(k_\nu)-1)} = (n_\nu^{(k_\nu)} - 1)\varphi + \varphi_\nu + \alpha + \frac{\pi}{2} - \theta_\nu^{(n-1)} < 2m_\nu\pi.$$

More generally, for $m = 0, 1, 2, \cdots$ we have

$$n_\nu^{(k_\nu+m)}\varphi + \varphi_\nu + \alpha + \frac{\pi}{2} - \theta_\nu^{(n)} \geqq 2(m_\nu + m)\pi$$

$$(n_\nu^{(k_\nu+m)} - 1)\varphi + \varphi_\nu + \alpha + \frac{\pi}{2} - \theta_\nu^{(n-1)} < 2(m_\nu + m)\pi.$$

We denote, for any real number a, by $[a]$ the largest integer not exceeding a. We also set

$$\psi_\nu = \varphi_\nu + \alpha + \frac{\pi}{2} + 2\pi(k_\nu - m_\nu). \tag{6}$$

If $k = k_\nu + m \geqq k_\nu$, it then follows that

$$n_\nu^{(k)} = \left[\frac{2\pi k - \psi_\nu + \theta_\nu^{(n)}}{\varphi}\right]. \tag{7}$$

For the proof of (iii) we observe that

$$n_\nu^{(k)} = \frac{2\pi k}{\varphi} + C_\nu^{(k)},$$

where the moduli of the numbers $C_\nu^{(k)}$ are bounded. We have

$$P_\nu^{(1)} + P_\nu^{(2)} + \cdots + P_\nu^{(k)} = n_\nu^{(k+1)} - n_\nu^{(1)}$$

and hence, using (7), if $k \geqq k_\nu$,

$$\frac{P_\nu^{(1)} + P_\nu^{(2)} + \cdots + P_\nu^{(k)}}{k} = \frac{2\pi}{\varphi} + \left(\frac{2\pi}{\varphi} + C_\nu^{(k)} - n_\nu^{(1)}\right)\frac{1}{k}.$$

The second term on the right tends to zero as $k \to \infty$, and (iii) follows.

For the proof of (iv) we set

$$N_\nu^{(k)} = \frac{1}{2}k - \frac{\varphi}{2\pi k}(n_\nu^{(1)} + n_\nu^{(2)} + \cdots + n_\nu^{(k)}).$$

The following lemma is required:

LEMMA. *If $\varphi/2\pi$ is irrational, then there exists a constant c such that for all ν satisfying $r_\nu \neq 0$*

$$\lim_{k\to\infty} N_\nu^{(k)} \equiv \frac{\varphi_\nu}{2\pi} + c \pmod 1.$$

The proof consists in showing that for a suitable integer l_ν and for every $\delta > 0$

(8a) $$\limsup_{k\to\infty} N_\nu^{(k)} \leqq \frac{\varphi_\nu}{2\pi} + c + l_\nu + \delta$$

and

(8b) $$\liminf_{k\to\infty} N_\nu^{(k)} \geqq \frac{\varphi_\nu}{2\pi} + c + l_\nu - \delta.$$

For $k \geqq k_\nu$, let

$$n_\nu^{(k)} = p_\nu^{(k)} + q_\nu^{(k)} + s_\nu^{(k)},$$

where

$$p_\nu^{(k)} = \frac{2\pi k - \psi_\nu}{\varphi},$$

$$q_\nu^{(k)} = \left[\frac{2\pi k - \psi_\nu}{\varphi}\right] - \frac{2\pi k - \varphi_\nu}{\varphi},$$

$$s_\nu^{(k)} = \left[\frac{2\pi k - \psi_\nu + \theta_\nu^{(n)}}{\varphi}\right] - \left[\frac{2\pi k - \psi_\nu}{\varphi}\right].$$

Let $\delta > 0$ be given, and let h be an integer such that

$$\frac{|\theta^{(n)}|}{\varphi} < \delta \quad \text{for} \quad k \geqq h,$$

where $n = n_\nu^{(k)}$. We then have

$$N_\nu^{(k)} = \frac{1}{2}k - \frac{1}{2\pi k}\left\{M + \varphi \sum_{m=h}^{k} (p_\nu^{(m)} + q_\nu^{(m)} + s_\nu^{(m)})\right\},$$

where

$$M = (n_\nu^{(1)} + n_\nu^{(2)} + \cdots + n_\nu^{(h-1)})\varphi.$$

Since M does not depend on k, $k^{-1}M \to 0$ as $k \to \infty$. An easy computation yields

$$\frac{1}{2}k - \frac{\varphi}{2\pi k}(p_\nu^{(h)} + p_\nu^{(h+1)} + \cdots + p_\nu^{(k)}) = -\frac{1}{2} + \frac{\psi_\nu}{2\pi} + \frac{1}{2\pi k}\{h(h-1) - (h-1)\psi_\nu\}.$$

The limit of this expression as $k \to \infty$ exists and equals $-\frac{1}{2} + \psi_\nu/2\pi$.

Since π/φ is irrational, the numbers $q_\nu^{(k)}$ are equidistributed in the interval $(-1, 0]$ according to a classical theorem by H. Weyl (see [6], p. 71, 234). Hence

$$\lim_{k\to\infty} \frac{1}{k}(q_\nu^{(h)} + q_\nu^{(h+1)} + \cdots + q_\nu^{(k)}) = -\frac{1}{2}.$$

According to the definition of h, the numbers $s_\nu^{(k)}$ can differ from 0 only if $q_\nu^{(k)}$ lies either in the interval $(-1, -1 + \delta)$ or in $(-\delta, 0]$. In either case, $|s_\nu^{(k)}| \leqq 1$. According to Weyl's theorem, the number of times either possibility occurs is asymptotic to $k\delta$ as $k \to \infty$. It follows that

$$\limsup_{k\to\infty} \frac{1}{k} |s_\nu^{(k)} + s_\nu^{(k+1)} + \cdots + s_\nu^{(k)}| \leqq \delta.$$

Gathering the above results, we find that

$$\limsup_{k\to\infty} N_\nu^{(k)} \leqq \frac{\psi_\nu}{2\pi} - \frac{1}{2} + \frac{\varphi}{2\pi}(-\tfrac{1}{2} + \delta),$$

$$\liminf_{k\to\infty} N_\nu^{(k)} \geqq \frac{\psi_\nu}{2\pi} - \frac{1}{2} + \frac{\varphi}{2\pi}(-\tfrac{1}{2} - \delta).$$

In view of (6), this establishes the relations (8) with

$$c = \frac{\alpha}{2\pi} - \frac{\varphi}{4\pi} - \frac{1}{4}, \qquad l_\nu = k_\nu - m_\nu.$$

The statement of the lemma now follows in view of the fact that the above is true for arbitrary $\delta > 0$.

Statement (iv) of the theorem now follows by observing the relation

$$\frac{\delta_{\nu\mu}^{(1)} + \delta_{\nu\mu}^{(2)} + \cdots + \delta_{\nu\mu}^{(k)}}{kP} = N_\mu^{(k)} - N_\nu^{(k)} \tag{9}$$

and letting $k \to \infty$.

If $2\pi/\varphi = q/p$ is rational and if $(p, q) = 1$, then the numbers $q_\nu^{(k)}$ are no longer equidistributed in $(-1, 0]$, but take on with equal frequency ([4], p. 51) the p distinct values

$$-\frac{m}{p} - \xi, \qquad m = 0, 1, \cdots, p - 1, \tag{10}$$

where ξ is some number depending on φ_ν, $0 \leqq \xi < 1/p$. It follows that

$$\lim_{k\to\infty} \frac{1}{k}(q_\nu^{(k)} + q_\nu^{(k+1)} + \cdots + q_\nu^{(k)}) = \frac{1 - p}{2p} - \xi. \tag{11}$$

If $\xi \neq 0$ in (10), the numbers $s_\nu^{(k)}$ are all zero for k sufficiently large, and thus

$$\lim_{k\to\infty} \frac{1}{k}(s_\nu^{(k)} + s_\nu^{(k+1)} + \cdots + s_\nu^{(k)}) = 0.$$

If $\xi = 0$, then $s_\nu^{(k)} = -1$ if $q_\nu^{(k)} = 0$ and $\theta_\nu^{(n)} < 0$, and $s_\nu^{(k)} = 0$ otherwise. We have $q_\nu^{(k)} = 0$ every pth time for k large, thus

$$\limsup_{k\to\infty} \frac{1}{k}(s_\nu^{(k)} + s_\nu^{(k+1)} + \cdots + s_\nu^{(k)}) \leqq 0,$$

$$\liminf_{k\to\infty} \frac{1}{k}(s_\nu^{(k)} + s_\nu^{(k+1)} + \cdots + s_\nu^{(k)}) \geqq -\frac{1}{p}.$$

Thus in any case, if $2\pi/\varphi$ is rational,

$$\limsup_{k\to\infty} N_\nu^{(k)} \leqq \frac{\psi_\nu}{2\pi} - \frac{1}{2} - \frac{\varphi}{4\pi}\left(1 - \frac{1}{p}\right),$$

$$\liminf_{k\to\infty} N_\nu^{(k)} \geqq \frac{\psi_\nu}{2\pi} - \frac{1}{2} - \frac{\varphi}{4\pi}\left(1 + \frac{1}{p}\right).$$

From these relations statement (v) of the theorem follows as above by observing (9) and using the relations

$$\limsup_{k\to\infty} \frac{\delta_{\nu\mu}^{(1)} + \delta_{\nu\mu}^{(2)} + \cdots + \delta_{\nu\mu}^{(k)}}{kP}$$

$$\leqq \limsup_{k\to\infty} N_\mu^{(k)} - \liminf_{k\to\infty} N_\nu^{(k)}$$

$$\leqq \frac{\psi_\mu - \psi_\nu}{2\pi} + \frac{\varphi}{2\pi p} \equiv \frac{\varphi_\mu - \varphi_\nu}{2\pi} + \frac{1}{q} \pmod 1$$

and a similar relation for the limit inferior.

5. Numerical Results.

1) As a basis for the numerical experiments we used the following 6×6 matrix A depending on two real parameters a and b, not both 0:

$$A = \alpha U D U^H, \quad \text{where}$$

$$\alpha = \frac{1}{|a| + |b|}$$

$$U = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ \frac{2}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{2}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} \\ \frac{-1}{\sqrt{12}} & \frac{2}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-2}{\sqrt{12}} \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & 0 & \frac{-2}{\sqrt{8}} & \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{2}{\sqrt{8}} & 0 & \frac{-1}{\sqrt{8}} \end{pmatrix}$$

$$D = \begin{Bmatrix} a-b & & b^2 & & & \\ -2 & & a+b & & & \bigcirc \\ & & & -4 & & \\ & & & & -2 & \\ & \bigcirc & & & & 1 \\ & & & & & 3 \end{Bmatrix}, \quad \text{and where}$$

U^H is the conjugate transpose of U.

The eigenvalues of A are seen to be:

$$\alpha(a + bi), \quad \alpha(a - bi), \quad -4\alpha, \quad -2\alpha, \quad \alpha, \text{ and } 3\alpha.$$

2) For $a = 6$, $b = \frac{1}{2}$, the resulting matrix A was as follows:

$$\begin{bmatrix} 0.2115385 & 0.3653846 & -0.1767767 & -0.05439283 & 0.3386381 & -0.1387861 \\ 0.3653846 & 0.2115385 & 0.1495803 & 0.3807498 & 0.02775722 & -0.005551444 \\ -0.05439283 & 0.2719641 & 0.2307692 & -0.3076923 & 0.2198260 & 0.5338631 \\ -0.1767767 & 0.2583659 & -0.3942308 & 0.5384615 & 0.05103104 & -0.2551552 \\ 0.3386381 & 0.02775722 & 0.1138384 & 0.1570186 & 0.2548077 & -0.1971154 \\ -0.03886011 & 0.09437457 & 0.5691923 & -0.2198260 & -0.1105769 & 0.09134615 \end{bmatrix}$$

For this matrix, the following results were obtained:

Eigenvalue

Computed Absolute Value	Computed Argument	Actual Absolute Value	Actual Argument
0.926276	0.083221 radians	0.926276	0.083140 radians

*Eigenvector**

Computed Absolute Value	Computed Argument	Actual Absolute Value	Actual Argument
1	33.45°	1	33.68°
1.00000	33.45°	1.00000	33.68°
0.392232	−90.00°	0.392232	−90.00°
0.980581	52.53°	0.980580	53.12°
0.866025	33.45°	0.866025	33.68°
0.537086	−115.84°	0.537086	−116.57°

3) Other cases:

a) For $a = \frac{1}{2}$ and $b = 5$, the results were as follows:

Eigenvalue

Computed Absolute Value	Computed Argument	Actual Absolute Value	Actual Argument
0.913625	1.46980 radians	0.913625	1.47113 radians

Eigenvector

Computed Absolute Value	Computed Argument	Actual Absolute Value	Actual Argument
1	0 radians	1	0 radians
1.00000	0.0000	1.00000	0.0000
1.05267	−0.2815	1.05267	−.2750
0.419137	2.377	0.419137	2.324
0.866025	0.0000	0.866025	0.0000
0.587022	−0.4378	0.587022	−0.4101

b) The case $a = 4$, $b = \frac{1}{2}$ proved to be of interest. Letting $\lambda_1, \cdots, \lambda_6$ denote the actual (theoretical) eigenvalues of the matrix corresponding to this case, it turns out that $|\lambda_1| = |\lambda_2| = 0.8958064$ and $|\lambda_3| = 0.8888888$, so that λ_3 is close

* Arguments in 2 were normalized so that the argument of the third component was −90°.

to λ_1 both in location and absolute value. The numerical process for finding the absolute value of λ_1 *did not converge* in this case, *but* the numerical procedure for finding the argument of λ_1 yielded 0.12433 radians as compared with the actual value of 0.12436 radians; i.e., the angle was obtained as accurately in this case as in cases where $|\lambda_1| = |\lambda_2| \gg |\lambda_3| > \cdots$.

c) The case $a = 0$, $b = 5$ yielded the eigenvalue (i.e., absolute value and argument) exactly.

NOTE: In each numerical example considered, (1,1,1,1,1,1) was used as the starting vector, and φ was determined from the average period of *all* components. The latter procedure was found to yield the angle φ more accurately than when the average period of just one component was used.

Space Technology Laboratories, Inc.
Los Angeles, California

University of California
Los Angeles, California

1. A. C. AITKEN, "Studies in practical mathematics, II. The evaluation of the latent roots and latent vectors of a matrix," *Proc. Roy. Soc. Edinburgh*, v. 57, 1937, p. 269–304.
2. E. BODEWIG, *Matrix Calculus*, North Holland Publishing Co., Amsterdam, 1959.
3. L. COLLATZ, *Eigenwertaufgaben mit Technischen Anwendungen*, Leipzig, 1949.
4. G. H. HARDY & E. M. WRIGHT, *An Introduction to the Theory of Numbers*, third edition, Oxford, 1954.
5. E. E. OSBORNE, "On acceleration and matrix deflation processes used with the power method," *J. Soc. Indust. Appl. Math.*, v. 6, 1958, p. 279–287.
6. G. POLYA & G. SZEGÖ, *Aufgaben und Lehrsätze der Analysis*, v. I., Springer, Berlin, 1954.
7. R. VON MISES & H. GEIRINGER, "Praktische Verfahren der Gleichungsauflösung," *Z. Angew. Math. Mech.*, v. 9, 1929, p. 58–77, p. 152–164.
8. H. WIELANDT, "Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben," *Math. Z.*, v. 50, 1954, p. 93–143.
9. J. H. WILKINSON, "The calculation of the latent roots and vectors of matrices on the Pilot model of the ACE," *Proc. Cambridge Philos. Soc.*, v. 50, 1954, p. 536–566.
10. J. H. WILKINSON, "The use of iterative methods for finding the latent roots and vectors of matrices," *MTAC*, v. 9, 1955, p. 184–191.

Acceleration Techniques for Iterated Vector and Matrix Problems*

By P. Wynn

1. Introduction. The ϵ-algorithm [1], which is closely related (see [2]) to the $e_m(S_n)$ transformation [3] and the method of summability by Stieltjes-type J-fractions [4], provides a powerful technique for transforming slowly convergent or divergent sequences (see for example [5], [6]). The quantities $\epsilon_s^{(m)}$ which satisfy the relationship

$$\epsilon_{s+1}^{(m)} = \epsilon_{s-1}^{(m+1)} + \frac{1}{\epsilon_s^{(m+1)} - \epsilon_s^{(m)}} \tag{1}$$

may be arranged in Table 1 and are seen to occur at the corners of a lozenge in this array.

The usefulness of the ϵ-algorithm lies in the fact that if the sequence

$$\epsilon_0^{(m)} = S_m \qquad m = 0, 1, \cdots . \tag{2}$$

is slowly convergent, then (in certain cases) the numerical convergence of the sequence $\epsilon_{2s}^{(0)} s = 0, 1, \cdots$ to the limit (or antilimit), with which the sequence (2) may be associated, is far more rapid.

In the application of the ϵ-algorithm so far the $\epsilon_s^{(m)}$ have been scalar quantities; it is the purpose of this paper to extend the inquiry to the cases in which the S_m are a sequence of slowly convergent arrays. In particular, the cases in which the S_m are (a) vectors (b) square matrices (c) triangular matrices will be considered.

The sums and differences of these entities are of course already well defined, but the choice of an inverse must be given some consideration. Four possibilities will be considered. They are

(1) *Primitive Inverse.* Regarding each component separately, this is equivalent to the simultaneous application of the scalar ϵ-algorithm to the components of (a), (b) and (c).

(2) *The Samelson Inverse of a Vector.* Here an extremely elegant and profound idea, due to K. Samelson, is introduced. The inverse of a vector

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

is taken to be

$$\mathbf{x}^{-1} = \left(\sum_{r=1}^{n} x_r \bar{x}_r\right)^{-1} (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n)$$

where $\bar{x}_r$ is the complex conjugate of x_r. The point $\mathbf{x}^{-1}$ is thus the inverse point of x with respect to the unit sphere in n-space.†

Received August 1, 1961.

* Communication MR of the Computation Department of the Mathematical Centre, Amsterdam.

† Dr. E. T. Goodwin has pointed out to the author that there is a strong connection between the idea of the Samelson inverse and a result ([22] p. 675, eq. (53)) of Lanczos.

TABLE 1

	$\epsilon_0^{(0)}$				
$\epsilon_{-1}^{(1)} = 0$		$\epsilon_1^{(0)}$			
	$\epsilon_0^{(1)}$		$\epsilon_2^{(0)}$		
$\epsilon_{-1}^{(2)} = 0$		$\epsilon_1^{(1)}$		$\epsilon_3^{(0)}$	
	$\epsilon_0^{(2)}$		$\epsilon_2^{(1)}$		$\epsilon_{3+1}^{(0)}$
$\epsilon_{-1}^{(3)} = 0$		$\epsilon_1^{(2)}$		$\epsilon_3^{(1)}$	
	$\epsilon_0^{(3)}$		$\epsilon_2^{(2)}$		$\epsilon_{3+1}^{(1)}$
				$\epsilon_3^{(2)}$	
					$\epsilon_{3+1}^{(2)}$

(3 and 4) *The Normally Defined Inverses of Square and Triangular Matrices.* (1) may be applied to all cases (a), (b) and (c); (2) may be applied to (b) and (c) in two ways, first by treating the rows (or columns) separately, and second by regarding the whole matrix as one vector.

From relationship (1) there may immediately be deduced:

THEOREM 1. *If a, S_m $\quad m = 0, 1, \cdots$ are of the same kind, b is a nonzero scalar and application of the ϵ-algorithm relationship to the initial values* (2) *produces quantities $\epsilon_s^{(m)}$, then application of the same relationship to the initial values*

$$\epsilon_{-1}^{(m)*} = 0 \qquad \epsilon_0^{(m)*} = a + bS_m \qquad m = 0, 1, \cdots \tag{5}$$

produces quantities

$$\epsilon_{2s}^{(m)*} = a + b\epsilon_{2s}^{(m)}, \qquad \epsilon_{2s+1}^{(m)*} = b^{-1}\epsilon_{2s+1}^{(m)} \qquad m, s = 0, 1, \cdots \tag{6}$$

Two results of F. L. Bauer [7] relating to the use of the primitive inverse may be extended to give

THEOREM 2

If in (2)

$$\sum_{s=0}^{h} c_s S_{m+s} = b \qquad m = 0, 1, \cdots \tag{7}$$

and the roots $\lambda_1, \lambda_2, \cdots, \lambda_h$ of the equation

$$\sum_{s=0}^{h} c_s \lambda^s = 0 \tag{8}$$

are real and distinct, and further

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_r| > 1 > |\lambda_{r+1}| > \cdots > |\lambda_h|$$

then

$$\lim_{m\to\infty} \epsilon_{2s-1}^{(m+1)} \epsilon_{2s-1}^{(m)-1} = \lambda_s^{-1} \tag{9}$$

and

$$\lim_{m\to\infty} \epsilon_{2s-2}^{(m+1)} \epsilon_{2s-2}^{(m)-1} = \begin{matrix} \lambda_s & s = 1, 2, \cdots, r \\ 1 & s = r+1, r+2, \cdots, h. \end{matrix} \tag{10}$$

The fundamental result in the theory of the ϵ-algorithm relating to scalar quantities is

THEOREM 3. *If equation* (7) *obtains and* $\sum_{s=0}^{k} c_s \neq 0$ *then*

$$\epsilon_{2k+1}^{(s)} = \infty, \qquad \epsilon_{2k}^{(s)} = b \Big/ \left(\sum_{i=0}^{k} c_i\right) \qquad s = 0, 1, \cdots. \tag{11}$$

This theorem is also true if the S_m have a suitably defined inverse and obey a non-commutative multiplication law. (It is quite clear, that if care is taken in the positioning of b two further versions of Theorem 1 may be given in which b is non-scalar). The proof of this result, however, is one of the more advanced results in the theory of continued fractions whose elements obey a non-commutative law of multiplication.

Before this proof can be described an almost complete account of this theory must be given, and this is far too extensive to be set down in detail here. The purpose of this paper is to demonstrate that the extension of the ϵ-algorithm to non-scalar quantities has useful application in numerical analysis, and to this we now proceed.

2. Iterated Vector Problems.

Sequential Relaxation. Perhaps the best known schemes for producing iterated vector sequences occur in the numerical solution of linear equations. Consider, for example, the process of obtaining the sequence $(x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)})$ $m = 1, 2, \cdots$ from a given initial vector $(x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)})$ by means of the scheme

$$\begin{aligned} a_{1,1}x_1^{(m+1)} + a_{1,2}x_2^{(m)} + \cdots + a_{1,n}x_n^{(m)} &= k_1 \\ a_{2,1}x_1^{(m)} + a_{2,2}x_2^{(m+1)} + \cdots + a_{2,n}x_n^{(m)} &= k_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n,1}x_1^{(m)} + a_{n,2}x_2^{(m)} + \cdots + a_{n,n}x_n^{(m+1)} &= k_n . \end{aligned} \tag{12}$$

If the primitive inverse is used it is to be expected that the conditions of Theorem 3 would apply and that the ϵ-algorithm would provide an n-step process for solving the set of equations

$$\mathbf{Ax} = \mathbf{k} \tag{13}$$

where $\mathbf{A}$, $\mathbf{x}$ and $\mathbf{k}$ are straight-forwardly defined by inspection of the quantities occurring in (12). (This is proved [8] by decomposing the vector $\mathbf{x}^{(0)}$ in terms of the eigenvectors of $\mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}$ is the main diagonal of $\mathbf{A}$). An example (Table 2, exponent in brackets), which gives the vector ϵ-scheme (primitive inverse) relating to the sequential relaxation of the set of equations

$$\begin{pmatrix} +2 & +1 & +3 & +4 \\ +1 & -3 & +1 & +5 \\ +3 & +1 & +6 & -2 \\ +4 & +5 & -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 4 \\ 8 \\ 6 \end{pmatrix} \tag{14}$$

TABLE 2

m	s								
	0	1	2	3	4	5	6	7	8
0	0.0								
	0.0								
	0.0								
	0.0	+2.0 (−1)							
		−7.5 (−1)							
		+7.5 (−1)							
1	+5.0	−1.666667 (−1)	−4.411765						
	−1.333333		+2.711864 (−1)						
	+1.333333		+3.168317 (−1)						
	−6.0	+9.375 (−2)	−2.16	+1.619173 (−1)					
		−1.267606 (−1)		−3.869420 (−1)					
		−2.337662 (−1)		−4.982868 (−1)					
2	+1.666667 (1)	+9.375 (−2)	+1.025802 (1)	+8.433735 (−2)	+4.085470				
	−9.222222		−3.572285		−1.340802				
	−2.944444		−3.463592		−9.252170 (−1)				
	+4.666667	−9.113924 (−2)	−1.084000 (2)	+4.330304 (−4)	−4.898484	+8.065247 (−2)			
		+5.023256 (−2)		+6.119052 (−2)		−7.528139 (−2)			
		−2.160000		−1.043340 (−1)		−2.813787 (−1)			
3	+4.694444	+8.490566 (−2)	+2.117836 (1)	+9.399904 (−2)	+1.653128 (1)	+1.533841 (−2)	−6.476506 (−1)		
	+1.068519 (1)		+8.768553 (1)		−8.668317		+1.605804		
	−3.407407		−2.977132		−6.573510		+6.325394 (−1)		
	+1.644444 (1)	−3.047404 (−2)	+1.570070	−2.147557 (−1)	−1.761132 (1)	+2.244161 (−2)	+1.243017	+4.332292 (−1)	
		+6.321951 (−2)		+5.081211 (−2)		+2.205053 (−2)		−1.032366	
		+1.640922 (−1)		−3.823916 (−1)		−1.426064 (−1)		+1.104180	
4	−2.812037 (1)	+1.767594 (−2)	+1.575188 (1)	+4.186519 (−2)	+2.074718 (1)	+6.837659 (−2)	+1.786698	−2.471836	+1.0
	+2.050309 (1)		+7.088480		−4.343692 (1)		+6.574128 (−1)		+1.0
	+2.686728		−4.807012		−2.403110		+1.434601		+1.0
	+7.301852 (1)	−7.680502 (−3)	+4.291075	−1.456719 (−2)	+2.010829 (1)	−3.029967 (−2)	+8.493496 (−1)	−8.379074 (−1)	+1.0
		+1.171190 (−2)		+3.102009 (−2)		+4.472918 (−2)		+1.886598	
		+3.064748 (−2)		+3.359876 (−2)		+1.179656 (−1)		−1.196780	
5	−1.583187 (2)	−1.553808 (−2)	−1.294576 (2)	−1.989735 (−3)	−4.281560 (1)	+1.645267 (−2)	+5.484726 (−1)	+4.166050	+1.0
	+1.118863 (2)		+5.887999 (1)		+2.950735 (1)		+1.200340		+1.0
	+3.531584 (1)		+3.340283 (2)		+0.449989		+6.739977 (−1)		+1.0
	+8.660494	+2.696816 (−2)	+1.167205 (2)	−3.025448 (−3)	+7.433117 (1)	−7.239109 (−3)	+1.090337	+1.376798	+1.0
		−7.153770 (−3)		−3.025208 (−3)		+9.402247 (−3)		−3.104928	
		+3.399518 (−2)		+3.051784 (−2)		+4.017075 (−3)		+1.870683	
6	−1.212379 (2)	−6.283961 (−3)	−1.627980 (2)	−2.558057 (−2)	−2.801389 (2)	+2.799081 (−3)	+1.270097	−6.903627	
	−2.790012 (1)		+3.010951 (2)		+1.099743 (2)		+8.792432 (−1)		
	+6.473178 (1)		+4.645171 (1)		−2.828486 (1)		+1.209712		
	−1.504748 (2)	+2.906646 (−3)	+6.499793 (1)	−1.154763 (−2)	+1.095677 (2)	−3.685574 (−3)	+9.455442 (−1)		
		−4.114212 (−3)		−8.257503 (−3)		+2.359331 (−4)			
		−2.070919 (−2)		+1.713751 (−2)		+3.792161 (−2)			
7	+2.228020 (2)	−1.640848 (−3)	−2.319818 (2)	−3.194056 (−3)	−1.529457 (2)	−6.407146 (−3)			
	−2.709600 (2)		+5.974104 (1)		+2.277123 (2)				
	+1.644404 (1)		+7.287409 (1)		+1.982883 (1)				
	−7.599158 (2)	+7.077924 (−4)	−5.789307 (2)	+1.104822 (−3)	−2.016593 (2)				
		−1.090333 (−3)		−2.304104 (−3)					
		−2.988134 (−3)		−1.714317 (−3)					
8	+1.635646 (3)	+3.884469 (−3)	+2.286721 (3)	−5.434449 (−4)					
	−1.188111 (3)		−7.641373 (2)						
	−3.183129 (2)		+8.579158 (2)						
	−5.024804 (2)	+2.243712 (−3)	−8.047707 (2)						
		+1.266304 (−3)							
		−2.137887 (−3)							
9	+2.081336 (3)	+5.763910 (−4)							
	−3.996564 (2)								
	−7.859644 (2)								
	+1.232453 (3)								

substantiates this. It will be observed that the results of Theorem 3 are not of much use in this case. But the situation is analogous to the process of determining the eigenvalues of a matrix by direct iteration, and Bodewig, to whom the matrix on the left-hand side of equation (14) is due, has pointed out [9] that this is a particularly unfavorable example.

Gauss-Seidel Relaxation. An alternative scheme for the iterative solution of equation (13) is provided by the scheme

$$
\begin{aligned}
a_{1,1}x_1^{(m+1)} + a_{1,2}x_2^{(m)} + \cdots + a_{1,n}x_n^{(m)} &= k_1 \\
a_{2,1}x_1^{(m+1)} + a_{2,2}x_2^{(m+1)} + \cdots + a_{2,n}x_n^{(m)} &= k_2 \\
a_{n,1}x_1^{(m+1)} + a_{n,2}x_2^{(m+1)} + \cdots + a_{n,n}x_n^{(m+1)} &= k_n .
\end{aligned} \tag{15}
$$

(When applied to the solution of the set (14), the results from using this scheme actually diverge even more wildly than those of (12).) Again, it is to be expected that the results of Theorem 3 apply (this time with $h = n - 1$).

The method is illustrated by the solution of

$$
\begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 9 \\ 6 & 8 & 10 & 7 \\ 5 & 7 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 23 \\ 32 \\ 33 \\ 31 \end{pmatrix} \tag{16}
$$

again using $\mathbf{x}_0^{(0)} = 0$. The results are shown (this time only the even-order columns of the ϵ-array are displayed) in Table 3, which refers to the use of the Samelson inverse. The results in Tables 2, 3, and 4 have been computed, as have all the results of this paper, using twelve-decimal floating point arithmetic and have been rounded off for presentation to seven decimals. It is indeed remarkable that use of the primitive inverse should serve to give reasonable estimates of three of the roots and yet be inadequate in giving the value of the fourth.

Some of the estimates (taking each root in isolation) produced by use of the primitive inverse are better than those produced by use of the Samelson inverse, but if a least squares indication of the error in the estimates is used, then use of the Samelson inverse is superior. This appears generally to be true.

The numerical results displayed in Tables 3 and 4 are not particularly impressive, but this is hardly to be expected since the Wilson matrix [10] (on the left-hand side of equation (16)) is well known to be ill-conditioned.

The foregoing results have served to bring certain points into focus, in particular that the user need not be unduly frightened by the rapid divergence of the sequence to be transformed, but in the event use of the ϵ-algorithm is to be recommended neither as a method for the direct solution of a set of linear equations nor as a technique for estimating the eigenvalues of a matrix.

Iterated vector sequences occur very frequently in optimization and approximation problems, but perhaps they occur most naturally in numerical analysis in the iterative solution by digital methods of equations in which the solution is a function of a single variable. More specifically, reference is being made to the solution of

$$
\alpha\{y(x), x\} = \beta\{y(x), x\} \tag{17}
$$

TABLE 3

m	s									
0	0									
	0									
	0									
	0									
1	+4.6		+3.810690							
	−2.0	(−2)	−1.766784	(−2)						
	+5.56	(−1)	+1.150607							
	+3.136	(−1)	+1.007187							
2	+3.647200		+2.262368		+2.276624					
	−1.736000	(−2)	+2.060921	(−2)	+2.016264	(−1)				
	+8.433280	(−1)	+1.091095		+6.618651	(−1)				
	+5.295568	(−1)	+1.050041		+1.198757					
3	+3.082754		+2.276609		+2.333036		+2.329996			
	−3.279680	(−3)	−5.669567	(−2)	+2.044079	(−1)	+9.996350	(−1)		
	+9.763705	(−1)	+1.047934		+6.628969	(−1)	+1.664600			
	+6.821856	(−1)	+1.085579		+1.198153		+9.997397	(−1)		
4	+2.750761		+2.287822		+2.329996		+9.999154	(−1)	+1.000277	
	+1.584073	(−2)	−2.504813	(−1)	+2.072094	(−1)	+1.000656		+1.000115	
	+1.022904		+1.022770		+6.639249	(−1)	+7.758344	(−1)	+9.039898	(−1)
	+7.929176	(−1)	+1.114529		+1.197555		+1.000214		+9.999536	(−1)
5	+2.557421		+2.296513		+2.327167		+1.000277		+1.000127	
	+3.644006	(−2)	+1.032570		+2.100183	(−1)	+9.995052	(−1)	+9.999178	(−1)
	+1.022769		+1.022904		+6.649406	(−1)	+3.329484	(−1)	+1.224386	
	+8.752889	(−1)	+1.137530		+1.196962		+9.996347	(−1)	+9.999873	(−1)
6	+2.446372		+2.303113		+2.324582		+1.000019			
	+5.662204		+3.382326	(−1)	+2.128077	(−1)	+1.000149			
	+9.991194	(−1)	+1.077096		+6.659566	(−1)	+8.660018	(−1)		
	+9.379713	(−1)	+1.155232		+1.196377		+1.000531			
7	+2.383815		+2.307990		+2.322279					
	+7.545437	(−2)	+2.610966	(−1)	+2.155269	(−1)				
	+9.651736		+1.389942		+6.669650	(−1)				
	+9.866184	(−1)	+1.168357		+1.195802					
8	+2.349537		+2.311459							
	+9.255223		+2.362798	(−1)						
	+9.282793	(−1)	−4.587012	(−1)						
	+1.024993		+1.177629							
9	+2.331498									
	+1.078324	(−1)								
	+8.923411	(−1)								
	+1.055661									

where α and β operate on $y(x)$ and x, by constructing the sequence $y_r(x) r = 0, 1, \cdots$ by means of the equation

$$\alpha\{y_{r+1}(x), x\} = \beta\{y_r(x), x\}. \tag{18}$$

If x ranges (or may be transformed from another variable so as to do so) from a to $a + nh$ then the iterated vectors become

$$(y_r(a), y_r(a + h), \cdots, y_r(a + nh)) \qquad r = 0, 1, \cdots$$

where h is a suitable interval. Examples of the application of the ϵ-algorithm in such cases now follow.

TABLE 4

m	s				
0	0				
	0				
	0				
	0				
1	+4.6	+3.780995			
	−2.0 (−2)	−1.748738 (−2)			
	+5.56 (−1)	+7.377991 (−2)			
	+3.136 (−1)	+4.537392 (−1)			
2	+3.647200	+2.290460	+2.310052		
	−1.736000 (−2)	+5.900807 (−2)	+7.761601 (−2)		
	+8.433280 (−1)	+1.036824	+9.780429 (−1)		
	+5.295568 (−1)	+9.803224 (−1)	+1.010201		
3	+3.082754	+2.310449	+2.327842	+2.323830	
	−3.279680 (−3)	+7.759300 (−2)	+2.011186 (−1)	+2.033882 (−1)	
	+9.763705 (−1)	+9.781162 (−1)	+6.644644 (−1)	+6.657814 (−1)	
	+6.821856 (−1)	+1.010156	+1.197278	+1.196510	
4	+2.750761	+2.329024	+2.323830	+1.000010	+1.000196
	+1.584073 (−2)	+9.340374 (−2)	+2.033883 (−1)	+1.000647	+9.995082 (−1)
	+1.022904	+9.287211 (−1)	+6.657814 (−1)	+9.994166 (−1)	+1.000566
	+7.929176 (−1)	+1.034256	+1.196510	+1.000180	+9.996452 (−1)
5	+2.557421	+2.347263	+2.319847	+1.000196	+1.000110
	+3.644006 (−2)	+1.071077 (−1)	+2.056189 (−1)	+9.995082 (−1)	+9.999540 (−1)
	+1.022769	+8.863379 (−1)	+6.671455 (−1)	+1.000566	+1.000010
	+8.752889 (−1)	+1.053689	+1.195713	+9.996452 (−1)	+1.000030
6	+2.446372	+2.366192	+2.315893	+1.000048	
	+5.662204 (−2)	+1.195974 (−1)	+2.078090 (−1)	+1.000149	
	+9.991194 (−1)	+8.482916 (−1)	+6.685656 (−1)	+9.995584 (−1)	
	+9.379713 (−1)	+1.069723	+1.194878	+1.000543	
7	+2.383815	+2.386159	+2.311967		
	+7.545437 (−2)	+1.323159 (−1)	+2.099600 (−1)		
	+9.651736 (−1)	+8.110905 (−1)	+6.700522 (−1)		
	+9.866184 (−1)	+1.084318	+1.193997		
8	+2.349537	+2.404774			
	+9.255223 (−2)	+1.475526 (−1)			
	+9.282783 (−1)	+7.706173 (−1)			
	+1.024993	+1.100770			
9	+2.331498				
	+1.078324 (−1)				
	+8.923411 (−1)				
	+1.055661				

Integral Equations: *Fredholm Integral Equations of the Second Kind.* An attempt to solve the equation

$$\int_a^b k(x,t)f(t)\,dt = g(x) + f(x) \tag{19}$$

may be made by setting up the classical iterative scheme

$$f_{r+1}(x) = -g(x) + \int_a^b k(x,t)f_r(t)\,dt. \tag{20}$$

An example of this technique is provided by the work of E. R. Love [11], who shows

that the equation

$$f(x) = 1 - \int_{-1}^{1} \frac{a}{\pi\{a^2 + (x-t)^2\}} f(t)\, dt \tag{21}$$

occurs in potential theory. This leads to the recursion among the sequence of continuous functions $f_r(x)$ $r = 0, 1, \cdots ; -1 \leqq x \leqq 1$

$$f_{r+1}(x) = 1 - \int_{-1}^{1} \frac{a}{\pi\{a^2 + (x-t)^2\}} f_r(t)\, dt. \tag{22}$$

Using the integration formulas

$$\int_a^{a+nh} f(t)\, dt = \tfrac{1}{2} f_0 + f_1 + f_2 + \cdots + f_{n-1} + \tfrac{1}{2} f_n + C \tag{23}$$

where

$$C = (-\tfrac{1}{12}\Delta + \tfrac{1}{24}\Delta^2 - \tfrac{19}{720}\Delta^3 + \cdots)(f_n - f_0) \tag{24}$$

or

$$C = \tfrac{1}{12}(\Delta f_0 - \nabla f_n) - \tfrac{1}{24}(\Delta^2 f_0 + \nabla^2 f_n) + \tfrac{19}{720}(\Delta^3 f_0 - \nabla^3 f_n) - \cdots \tag{25}$$

(22) is transformed into an iteration scheme among the vectors

$$(f_r(-1), f_r(-1+h), \cdots, f_r(1-h), f_r(1)) \qquad r = 0, 1, \cdots .$$

Table 5 shows the results of this scheme and the effects of accelerating it (Samelson inverse), when $a = 1.0$ and $h = 0.25$.

Due to the symmetry of the interval of integration and the fact that the kernel is a function of $x - y$, the solution vector is of course symmetric about the origin, and for this reason the values of the vectors for negative argument are not shown. The transformed results check to four decimals those derived by Fox and Goodwin [12] by use of the more normal method of approximating (21) by a set of linear equations. (These authors also recommend, with the support of an example, the acceleration of the iterative scheme (22) by means of Aitkens' δ^2 process [13], which corresponds in the present notation to the use of the sequence $\epsilon_2^{(s)}$ $s = 0, 1, \cdots$.)

Volterra Integral Equations. The equations now being considered, are of the form

$$\int_a^x k(x,t) f(t)\, dt = g(x) + f(x) \tag{26}$$

and lead to the iterative scheme

$$f_{r+1}(x) = -g(x) + \int_a^x k(x,t) f_r(t)\, dt. \tag{27}$$

An example of such a scheme is provided by the integral equation

$$\int_0^x \frac{-2}{(x-t+2)^2} f(t) = f(x) - \frac{1}{(x+2)^2} \tag{28}$$

of Friedlander [14] occurring in the theory of parabolic reflectors. This leads to the

TABLE 5

m	s			
	0	2	4	
0	1.0			
	1.0			
	1.0			
	1.0			
	1.0			
1	0.49999	0.65702		
	0.50990	0.66353		
	0.53961	0.68323		
	0.58731	0.71524		
	0.64757	0.75623		
				x
2	0.72889	0.65743	0.65741	0.00
	0.73367	0.66382	0.66381	0.25
	0.74833	0.68320	0.68320	0.50
	0.77271	0.71489	0.71490	0.75
	0.80465	0.75568	0.75570	1.00
3	0.62499	0.65741		
	0.63212	0.66381		
	0.65366	0.68320		
	0.68868	0.71490		
	0.73350	0.75570		
4	0.67212			
	0.67818			
	0.69660			
	0.72679			
	0.76577			

recursion

$$f_{r+1}(x) = (x+2)^{-2} - 2\int_0^x (x-t+2)^{-2} f_r(t)\, dt. \tag{29}$$

Again, by the use of formula (23) this is transformed into an iterated vector scheme. Table 6 gives the results of this scheme and the accelerated results (primitive inverse). The integration is carried out over the range of $0 \leqq x \leqq 1.75$ at an interval in x of 0.25. After a further two iterations of (29), the original sequence would have attained the accuracy of the transformed results shown and, in view of the simplicity of (29) and (23), this means that this example is not a particularly striking one; however use of the acceleration technique here is instructive, and if the original equation had been more complicated (as in practice such equations normally are) it would have been useful.

The results may be checked by use of the power series expansion

$$f(x) = r_0 - \tfrac{1}{2}xr_1 + (\tfrac{1}{2}x)^2 r_2 - \cdots \tag{30}$$

where

$$\tau_0 = 1,$$

$$\tau_n = \frac{1}{4}(n+1) + \tau_0 + \frac{\tau_1}{\binom{n}{1}} + \frac{\tau_2}{\binom{n}{2}} + \cdots + \frac{\tau_{n-1}}{\binom{n}{n-1}}, \quad n = 1, 2, \cdots \tag{31}$$

derived by Fox and Goodwin [12] directly from the integral equation (28). For

TABLE 6

m	s			
	0	2	4	
0	0.25			
	0.197 531			
	0.160 000			
	0.132 231			
	0.111 111			
	0.094 675			
	0.081 633			
1	0.25	0.25		
	0.172 960	0.174 396		
	0.120 665	0.124 994		
	0.084 293	0.091 740		
	0.058 491	0.068 767		
	0.039 921	0.052 535		
	0.026 386	0.040 828		
				Correct
2	0.25	0.25	0.25	0.25
	0.174 485	0.174 427	0.174 425	0.174 424
	0.125 530	0.125 161	0.125 154	0.125 152
	0.093 110	0.092 145	0.092 119	0.092 118
	0.071 260	0.069 466	0.069 408	0.069 411
	0.056 310	0.053 549	0.053 446	0.053 444
	0.045 939	0.042 144	0.041 987	0.041 980
3	0.25	0.25		
	0.174 424	0.174 425		
	0.125 131	0.125 153		
	0.092 026	0.092 116		
	0.069 172	0.069 401		
	0.052 989	0.053 430		
	0.041 231	0.041 957		
4	0.25			
	0.174 425			
	0.125 154			
	0.092 125			
	0.069 429			
	0.053 498			
	0.042 090			

large values of x this series is unsuitable for direct use, but may be transformed into a continued fraction (again using the ϵ-algorithm), which converges quite rapidly.

Ordinary Differential Equations. The operational equation (17) may also refer to a differential equation. For example, the differential equation

$$y'' + f(x, y) = 0 \tag{32}$$

may be transcribed [15] as

$$\begin{aligned} y(x+h) + \phi\{x+h, y(x+h)\} = \\ 2y(x) - 10\phi\{x, y(x)\} - y(x-h) - \phi\{x-h, y(x-h)\} + \Delta\{y(x)\} \end{aligned} \tag{33}$$

where

$$\phi\{x, y(x)\} = \frac{h^2}{12} f(x, y), \tag{34}$$

h is the interval of tabulation, and

$$\Delta = 0\{\delta^6 y(x)\}. \tag{35}$$

Neglecting Δ, the following iteration scheme may be derived

$$\begin{aligned} y_{r+1}(x+h) - 2y_{r+1}(x) + y_{r+1}(x-h) \\ = -\phi\{x+h, y_r(x+h)\} - 10\phi\{x, y_r(x)\} - \phi\{x-h, y_r(x-h)\}. \end{aligned} \tag{36}$$

This may be solved for $y_{r+1}(x+h)$ to give a forward recursion scheme using values of $y(x)$ computed independently when $x = a, a+h$, or used as the basis for a relaxation scheme using a two-point boundary condition giving values of $y(x)$ at $x = a$ and $x = a + (n+1)h$. The latter has been carried out for the equation [15]

$$y'' - \frac{x^2 + 1/4}{x^2} y - \frac{\pi^2}{y^3} = 0 \tag{37}$$

which is satisfied by

$$y = \pi\{x(J_0(x)^2 + Y_0(x)^2)/2\}^{1/2} \tag{38}$$

for $a = 5.0$, $h = 0.5$, $n = 5$. The results are displayed together with the accelerated results (Samelson inverse) in Table 7. $y_0(x)$ is derived from equation (36) by replacing the right-hand side by zero.

The value of $y(5.0)$ has been computed with the assistance of the power series expansions

$$\begin{aligned} J_0(x) &= \sum_{r=0}^{\infty} \frac{(-1)^r}{(r!)^2} \left(\frac{x}{2}\right)^{2r} \\ Y_0(x) &= 2\left\{\sum_{r=0}^{\infty} \frac{(-1)^r}{(r!)^2} \left(\frac{x}{2}\right)^{2r} \left(\log\left(\frac{x}{2}\right) - \psi(r)\right)\right\}/\pi. \end{aligned} \tag{39}$$

and the value of $y(8.5)$ by use of the formula

$$y(x) = \{\pi(P_0(x)^2 + Q_0(x)^2)\}^{1/2} \tag{40}$$

where

$$P_0(x) \sim 1 - \frac{1^2 3^2}{2!(8x)^2} + \frac{1^2 3^2 5^2 7^2}{4!(8x)^4} - \frac{1^2 3^2 5^2 7^2 9^2 11^2}{6!(8x)^6} + \cdots$$

(41)

$$Q_0(x) \sim \frac{-1^2}{1!(8x)} + \frac{1^2 3^2 5^2}{3!(8x)^3} - \frac{1^2 3^2 5^2 7^2 9^2}{5!(8x)^5} + \cdots$$

(In the event these asymptotic series are not of much use for the argument given, but thirteen terms of the corresponding continued fraction expansion gives twelve-decimal accuracy).

Inspection of Table 7 reveals that both the original and the transformed results are better at the end than at the beginning of the range of x; this merely reflects

TABLE 7

m	s			
	0	2	4	
0	1.76865			
	1.76904			
	1.76942			
	1.76980			
	1.77019			
	1.77057			
1	1.76953	1.76880		
	1.77066	1.76931		
	1.77144	1.76976		
	1.77180	1.77014		
	1.77177	1.77045		
	1.77144	1.77071		
				Correct
2	1.76509	1.76880	1.76880	1.76897
	1.76264	1.76932	1.76931	1.76950
	1.76145	1.76977	1.76976	1.76992
	1.76184	1.77014	1.77014	1.77026
	1.76380	1.77045	1.77045	1.77054
	1.76703	1.77071	1.77071	1.77076
3	1.78737	1.76888		
	1.80279	1.76948		
	1.81149	1.76998		
	1.81184	1.77035		
	1.80386	1.77062		
	1.78924	1.77080		
4	1.67863			
	1.60708			
	1.56771			
	1.56816			
	1.60839			
	1.68069			

the fact that in (35) the differences increase as the origin is approached. For an alternative iterative scheme based upon a forward recursion see [15].

In equation (33) Δ is, of course, the difference correction of Fox and Goodwin's method VII [16]. The divergence of successive iterates is however so wild that it is futile to apply this technique. The variation in $y_r(x)$ with r is due more to the variation of $\phi\{x, y_r(x)\}$ with r than to the variation of the difference correction.

A case in which the variation in the iterates is due solely to the variation in the difference correction is provided by the equation [17]

$$y'' - (z^{-4} - 2z^{-3})y + z^{-2} = 0 \tag{42}$$

which is satisfied by the function

$$T(z) \sim 1!z^2 + 2!z^3 + 3!z^4 + \cdots . \tag{43}$$

Equation (42) serves as a basis for the recursion scheme

$$\begin{aligned}&(1 - h^2f(z + h)/12)y_{r+1}(z + h)\\ &- (2 + 5h^2f(z)/12)y_{r+1}(z) + (1 - h^2f(z - h)/12)y_{r+1}(z - h)\\ &= -h^2\{5(6z^2)^{-1} + (12(z + h)^2)^{-1} + (12(z - h)^2)^{-1}\} - \Delta\{y_r(z)\}\end{aligned} \tag{44}$$

where

$$f(z) = 2z^{-3} - z^{-4}$$

and

$$\Delta\{y(z)\} \equiv \{{}^{6}_{h}/240 - 13{}^{8}_{h}/15120 + 17{}^{10}_{h}/10080 - {}^{12}_{h}/29700\}y(z). \tag{45}$$

Values of $T(z)$ for small negative values of z may be computed by using the continued fraction expansion

$$T(-z) = \frac{z^2}{1+}\,\frac{2z}{1+}\,\frac{1z}{1+}\,\frac{3z}{1+}\,\frac{2z}{1+}\,\frac{4z}{1+}\,\frac{3z}{1+}\,\frac{\left[\frac{r}{2}\right]z}{1+}\,\frac{\left(\left[\frac{r}{2}\right]+1\right)z}{1+} \tag{46}$$

while for larger values of z the power series expansion

$$T(-z) = x^{-1} + e^x\left\{\gamma + \log(x) - \sum_{n=1} \frac{(-1)^{n-1}x^n}{n(n!)}\right\}, \qquad x = z^{-1} \tag{47}$$

is available.

Using these expansions to provide boundary conditions at $z = -0.5$ and $z = -1.2$, and further function values outside these points to assist in the computation of (45), (44) has been used with $h = 0.1$ to produce a vector sequence, which has been accelerated using the primitive inverse (see Table 8).

It is, of course, a feature of the iterated solution of two-point boundary condition problems that the iterates are available as vectors, but if forward recursion from one point is used, each function value may be iterated and corrected before proceeding to the next, and this is of course to be preferred.

3. Iterated Matrix Problems. As in the case of iterated vector problems, matrix sequences occur most naturally in numerical analysis in the digital solution of

TABLE 8

m	ε		
	0	2	4
0	0.18526 42708 0.23554 25892 0.28896 68980 0.34511 58040 0.40365 25004 0.46430 26898		
1	0.18526 44914 0.23554 28499 0.28896 71560 0.34511 60219 0.40365 26582 0.46430 27789	0.18526 44677 0.23554 28488 0.28896 71465 0.34511 60156 0.40365 26541 0.46430 27704	
2	0.18526 44649 0.23554 28488 0.28896 71462 0.34511 60154 0.40365 26540 0.46430 27695	0.18526 44705 0.23554 28510 0.28896 71490 0.34511 60161 0.40365 26545 0.46430 27714	0.18526 44702 0.23554 28473 0.28896 71489 0.34511 60164 0.40365 26543 0.46430 27714
3	0.18526 44720 0.23554 28465 0.28896 71501 0.34511 60162 0.40365 26545 0.46430 27719	0.18526 44702 0.23554 28473 0.28896 71489 0.34511 60164 0.40365 26543 0.46430 27714	
4	0.18526 44696 0.23554 28477 0.28896 71483 0.34511 60164 0.40365 26542 0.46430 27713		

iterative problems involving functions of two variables. A most fruitful source of such problems lies in the theory of partial differential equations, and examples of their treatment will now be given.

Boundary Value Problems. Two examples of the acceleration of the iterative solution of boundary value problems will be given. They are both provided by the equation

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} - \frac{2\partial^2\phi}{\partial x \partial y} = 0. \tag{48}$$

This, with the appropriate boundary conditions, has the solution

$$\phi = x/(x + y).$$

In both cases the boundary is the square whose vertices in the x, y plane are at

(49) $\quad 1, 1 + (n+1)h; \quad 1 + (n+1)h, 1 + (n+1)h; \quad 1 + (n+1)h, 1; \quad 1, 1.$

The derivatives in (48) have been approximated by means of the well-known schemes [18]

$$(50) \qquad \begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix} \phi_{x,y} \doteq 6h^2 \left\{ \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} \right\}, \quad \begin{bmatrix} -1 & & 1 \\ & 0 & \\ 1 & & -1 \end{bmatrix} \phi_{x,y} \doteq 4h^2 \frac{\partial^2 \phi}{\partial x \partial y}.$$

In the first example equation (48) has been rearranged to give the iteration scheme

$$(51) \qquad \frac{\partial^2 \phi_{r+1}}{\partial x^2} + \frac{\partial^2 \phi_{r+1}}{\partial y^2} = \frac{2\partial^2 \phi_r}{\partial x \partial y}.$$

Use of (50) gives a system of linear equations for the n^2 function values at the lattice points interior to the square (49). The right-hand sides of these equations are estimated by the use of (50) with modification of the appropriate equations by means of the boundary conditions. The first member of the sequence $\phi_0(x, y)$ is obtained by replacing the right-hand side of equation (51) by zero. The results of this iterative scheme (when $h = 0.25$ and $n = 4$) and subsequent acceleration using a matrix inverse are displayed in Table 9. (It is neither desirable nor feasible to display the complete matrices in this case. Instead the scalar norm $n^{-2} \sum_{i,j} | \phi_r(x, y) - x/(x + y) |$ has been given. This is so for the other tables in this section).

Equation (48) has also been used to derive the scheme

$$(52) \qquad \frac{\partial^2 \phi_{r+1}}{\partial x \partial y} = \frac{1}{2} \left\{ \frac{\partial^2 \phi_r}{\partial x^2} + \frac{\partial^2 \phi_r}{\partial y^2} \right\}.$$

The treatment here is precisely as in the preceding case. (It is perhaps of interest to remark that it is easy to construct the inverse of the matrix occurring in the resulting set of linear equations; this means that multiplication by this inverse matrix may be carried out analytically, and there is no need, as in the preceding case, to solve the set of equations numerically (see [23])).

The results of the iteration procedure and subsequent acceleration using the Samelson inverse vector-wise, when $h = 0.1$, $n = 4$, are shown in Table 10. The behavior of the iterated solutions is most odd in this example. The surfaces flap

TABLE 9

m	s		
	0	2	4
0	0.71 (−3)		
1	0.18 (−3)	0.93 (−4)	
2	0.65 (−4)	0.13 (−4)	0.61 (−5)
3	0.37 (−4)	0.70 (−5)	
4	0.11 (−4)		

TABLE 10

m	s					
	0	2	4	6	8	10
0	.62 (−3)					
1	.77 (−2)	.41 (−3)				
2	.15 (0)	.24 (−2)	.40 (−3)			
3	.29 (+1)	.12 (−1)	.13 (−2)	.53 (−4)		
4	.56 (+2)	.96 (−1)	.25 (−1)	.11 (−3)	.90 (−5)	
5	.11 (+4)	.52 (0)	.56 (−1)	.21 (−3)	.21 (−4)	.34(−5)
6	.21 (+5)	.40 (+1)	.13 (+2)	.70 (−3)	.12 (−3)	
7	.40 (+6)	.23 (+2)	.23 (+1)	.19 (−2)		
8	.77 (+7)	.17 (+3)	.37 (+2)			
9	.15 (+9)	.99 (+3)				
10	.28 (+10)					

about the line $x = y$, $\phi = 0$ with increasing amplitude. This feature is preserved during the accelerations.

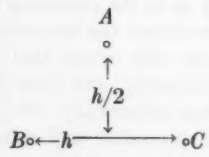
Initial Value Problems. Here the equation which has been selected for treatment is

$$\phi_{xx} - \phi_{yy} = 4\phi^2. \tag{53}$$

Initial values of ϕ, $\partial\phi/\partial x$ and $\partial\phi/\partial y$ have been given at the points $1, 0; 1 + h, 0; \cdots; 1 + (n + 1)h, 0$. The solution to this equation with appropriate initial values is

$$\phi = (x^2 - y^2)^{-1}. \tag{54}$$

If A, B and C are three neighboring points and BA and CA are characteristics of equations (53)



then the numerical integration of (35) proceeds according to the schemes (see e.g., (19))

$$\begin{aligned} U_A &= (U_C + V_C - U_B - V_B)/2 - h(M_C - M_B)/8 \\ V_A &= (U_C + V_C - U_B + V_B)/2 - h(2M_A + M_C + M_B)/8 \\ \phi_A &= (\phi_B + \phi_C)/2 + h(U_B + V_B + V_C - U_C + 2V_A)/8 \end{aligned} \tag{55}$$

where

$$U \equiv \frac{\partial\phi}{\partial x}, \qquad V \equiv \frac{\partial\phi}{\partial y} \quad \text{and} \quad M \equiv 4\phi^2.$$

Thus with the boundary conditions stated, a sequence of triangular arrays are produced. The first member of the sequence was obtained by replacing the right-

hand side of (53) by zero. The results of this process together with the accelerated results (primitive inverse) when $h = 0.25$, $n = 12$ are shown in Table 11.

This example has been chosen chiefly to show that the method proposed is feasible. It would seem more economical to use here a special feature of two-dimensional initial value problems, namely, that successive iterates are available as vectors rather than as matrices as in the case of boundary value problems. This implies a reduction in storage space for the acceleration technique; also that if the iteration itself converges rapidly at any stage then this advantage is preserved and the acceleration technique held in the reserve to be used when necessary.

4. Note on Programming. When the quantities involved are scalar the ϵ-algorithm may be put into effect in the following way. Passing from left to right in Table 1, each column is displaced downward half a row with respect to its left-hand neighbor; the ϵ-array is thus transformed into half a square array (the e-array, say). The quantities $\epsilon_s^{(m)}$ $m = 0, 1, \cdots, m' - s$ have now become $e_{i,j}$ $j = s$ $i = j, j+1, \cdots, m'$. Assuming that the initial sequence $S_m = \epsilon_0^{(m)} = e_0^{(m)}$ $m = 0, 1, \cdots, m'$ has been inserted, the construction of the e-array then runs

$$\text{for } j := 1(1)m' \text{ do for } i := j(1)m' \text{ do}$$

$$\text{if } j = 1 \text{ then } e_{i,1} := (e_{i,0} - e_{i-1,0})^{-1}$$

$$\text{else } e_{i,j} := e_{i-1,j-2} + (e_{i,j-1} - e_{i-1,j-1})^{-1}.$$

This implies a prodigal use of the storage space but the "program" is quite simple.

When the quantities involved are vectors or matrices the implied wastage of storage space in the preceding procedure may be unacceptable. In that event it is only necessary to retain a linear array during the formation of the ϵ-scheme.

This line of vectors or matrices corresponds to what would in a normal difference table be referred to as a line of backward differences; it contains the quantities $(\epsilon_0^{(m)}, \epsilon_1^{(m-1)}, \epsilon_2^{(m-2)}, \cdots, \epsilon_m^{(0)}$. After computing a new line these quantities are sequentially replaced by the quantities $\epsilon_0^{(m+1)}, \epsilon_1^{(m)}, \epsilon_2^{(m-1)}, \cdots, \epsilon_{m+1}^{(0)}$, incidentally increasing the length of the line by one unit. It is necessary to retain two extra blocks of intermediate storage space during the computation which proceeds as

TABLE 11

m	s			
	0	2	4	6
0	.54 (−1)			
1	.11 (−1)	.13 (−2)		
2	.99 (−3)	.78 (−4)	.97 (−4)	
3	.17 (−3)	.98 (−4)	.97 (−4)	.97(−4)
4	.93 (−4)	.97 (−4)	.97 (−4)	
5	.97 (−4)	.97 (−4)		
6	.97 (−4)			

follows:

compute $\epsilon_0^{(0)}$; $*\ l_0 = \epsilon_0^{(0)}$; $m := 1$;

EPSALG: compute $\epsilon_0^{(m)}$; $s := 0$; $b := \epsilon_0^{(m)}$;

EPSLOOP: recip := inverse of $(b - l_s)$;

‡ if $s \neq 0$ then begin *recip* := *recip* + l_{s-1}; $l_{s-1} := c$ end;

$c := b$; $b :=$ recip; $\overset{*}{*}$ $s := s + 1$; if $m - 1 \geqq s$ then go to EPSLOOP;

$l_{m-1} := c$; $l_m := b$; †$m := m + 1$; if $m_{\max} \geqq m$ then go to EPSALG.

There are a number of comments which should be made. First, the value of m in this program is the value of m in $\epsilon_0^{(m)}$ which is at the beginning of the line l. Second, the meaning of the := sign depends upon the nature of the quantities involved. Thus, in the case of n-length vectors "$c := b$" should be interpreted as "for $i := 1(1)n$ *do* $c_i := b_i$". Third, if the quantities $\epsilon_0^{(m+1)} - \epsilon_0^{(m)}$ are being computed, rather than $\epsilon_0^{(m)}$, then it may improve stability to substitute

EPSALG: compute $(\epsilon_0^{(m+1)} - \epsilon_0^{(m)})$; $\quad b := l_0 + (\epsilon_0^{(m+1)} - \epsilon_0^{(m)})$;

recip := inverse of $(\epsilon_0^{(m+1)} - \epsilon_0^{(m)})$;

set s to zero, and enter at ‡, rather than at EPSLOOP. Fourth, in any provisional inquiry it is useful to sample in some way the quantities $\epsilon_s^{(m)}$ with even suffices (as in Tables 9 to 11). This can be done by inserting the following instructions at the points * and $\overset{*}{*}$:

$*$ *display*$_0$:= *norm* $(\epsilon_0^{(0)})$;

$\overset{*}{*}$ if $s = 0$ mod (2) then *display* $_{s(2m_{\max}-s)/4+m}$:= *norm* (c);

if $s = m - 1$ and $m = 0$ mod (2) then *display* $_{m(2m_{\max}-m+4)/4}$:= *norm* (b);

This maps scalar norms of the quantities $\epsilon_s^{(m)}$ with even suffix upon a linear array *display*$_i$ $i = 0(1)([m_{\max}/2] + 1)^2$. They may be extracted after completion of the ϵ-scheme computation by means of the following program:

for *sanfang* := $0(2col)m_{\max}$ do begin New Line Carriage Return;

for $m := 0(1)m_{\max} -$ *sanfang* do begin New Line Carriage Return;

for $s :=$ *sanfang* (2) *sanfang* + 2*col*-2 do if $[s/2] \leqq m \leqq m_{\max} - [s/2]$

then print (*display*$_{s(2m_{\max}+2-s)/4+m}$) end m end *sanfang*;

This prints out a table having the format of Tables 9 to 11 in strips of *col* columns which can subsequently be glued into position. If a non-scalar sample (of $n + 1$ quantities, say) is required, then the insertions at * and $\overset{*}{*}$ must suitably be ex-

tended and the printing out must be done as follows:

for $s := 0(2)\ m_{max}$ do begin New Line Carriage Return;

for $m := 0(1)\ m_{max} - s$ do for $i := 0(1)\ n$ do begin

if $i = 0 \bmod(col)$ then New Line Carriage Return;

print $(sample_{i,s(2m_{max}+2-s)/4+m})$ end i end s.

Fifth, reference must be made to criteria for terminating the ϵ-process. (In the above program it has been assumed that m_{max} is known.) In the case of the scalar ϵ-process, the author's experience with a large number of examples has been that if the ϵ-process converges to a limit L and $|\ \epsilon_m^{(0)} - \epsilon_{m-2}^{(2)}\ | \leqq \delta$ for some even m then $|\ L - \epsilon_m^{(0)}\ | \leqq \delta$. This is a purely empirical test and one can easily construct examples in which it breaks down (indeed, one has a singly infinite number of degrees of freedom with which to do so). Nevertheless, if desired, this test may be inserted at the point † and carried out as follows † if $m = 0 \bmod(2)$, then if $|\ b - l_{m-2}\ | \leqq \delta$, go to EXIT.

At the point in the program labelled EXIT the required limit is, of course, b.

It is essential to stress that this type of estimation must be applied with discretion in the case of vector and matrix sequences.

An example (Table 12) will make this clear. This refers to the iteration of the scheme (44) using the first term of the difference correction and its subsequent acceleration (primitive inverse) when $h = 0.5$. To save space the successive estimates of $T(1.5)$ have been extracted and shown in a scalar ϵ-scheme. The same sort of behavior is shown by all components of the iterated vectors.

Judging from the table, one might be excused for thinking that the last column contains the required result correct to 11 decimals; but this is not at all true (the agreement is good only to 4 decimals). It is essential to obtain an independent estimate of how good the results are likely to be (in most cases this will be provided by examining remainder terms needed in the finite difference formulas. This applies in all cases of this paper). The terms

$$-\delta_{0.5}^{6} y_r(1.5)/240 \qquad \text{for} \qquad r = 0(1)3 \text{ are}$$

−0.0001 2325 120, −0.0001 1909 511, −0.0001 1925 940, −0.0001 4925 181, indeed provide slightly pessimistic estimates of the truncation errors in the original sequence and in the final transformed result.

TABLE 12

m	s		
	0	2	4
0	0.7239 5826 349		
1	0.7240 5913 419	0.7240 5713 559	
2	0.7240 5709 519	0.7240 5716 053	0.7240 5716 006
3	0.7240 5716 270	0.7240 5716 005	
4	0.7240 5715 994		

The point to be made here is that the original sequence undoubtedly converges to some limit. The ϵ-algorithm in this case is a convergence-preserving device. However, in view of the large interval of tabulation, this limit and the transformed result are not very interesting.

5. Conclusion. It is perhaps of interest to comment on reasons for preferring the use of one inverse in the ϵ-algorithm to any other. The consistently good numerical agreement between results produced by using the various inverses in any one case leads the author to conjecture that, within a certain limit, the rate of convergence of the ϵ-process is independent of the inverse. This being so, the choice of an inverse will depend on relative complexity and on numerical stability. On both counts the matrix inverses can be discarded. This leaves the choice between the primitive and Samelson inverses, and the author would suggest the latter for the following reason. If in any application of the ϵ-algorithm $\epsilon_{s-2}^{(m+1)} = \epsilon_{s-2}^{(m+2)} = \alpha$ then the following situation obtains:

$$
\begin{array}{ccccc}
 & & \epsilon_{s-1}^{(m)} & & \\
 & (\alpha)\epsilon_{s-2}^{(m+1)} & & (\alpha)\epsilon_{s}^{(m)} & \\
\epsilon_{s-3}^{(m+2)} & & (\infty)\epsilon_{s-1}^{(m+1)} & & (?)\epsilon_{s+1}^{(m)} \\
 & (\alpha)\epsilon_{s-2}^{(m+2)} & & (\alpha)_{s}\epsilon^{(m+1)} & \\
 & & \epsilon_{s-1}^{(m+2)} & &
\end{array}
\tag{56}
$$

the value of $\epsilon_{s+1}^{(m)}$ computed by means of the ϵ-algorithm becomes indeterminate. This is by no means an exceptional occurrence. It can be overcome by elimination and one obtains the singular rules

$$\epsilon_{s+1}^{(m)} = \epsilon_{s-1}^{(m+2)} + \epsilon_{s-1}^{(m)} - \epsilon_{s-2}^{(m+2)} \tag{57}$$

when

$$\epsilon_{s-2}^{(m+1)} = \epsilon_{s-2}^{(m+2)}$$

and

$$\epsilon_{s+1}^{(m)} = a(1 + a\epsilon_{s-1}^{(m+1)-1})^{-1} \tag{58}$$

where

$$a = \epsilon_{s-1}^{(m+2)}(1 - \epsilon_{s-1}^{(m+2)}\epsilon_{s-1}^{(m+1)-1})^{-1} + \epsilon_{s-1}^{(m)}(1 - \epsilon_{s-1}^{(m)}\epsilon_{s-1}^{(m+1)-1})^{-1} \tag{59}$$

should $\epsilon_{s-2}^{(m+1)}$ and $\epsilon_{s-2}^{(m+2)}$ be nearly equal. But this complicates the program a great deal. (The misfortune just described is of course recognized as a state of chronic instability, for example see Table 3.) At least, when using the Samelson inverse the probability of encountering this misfortune is reduced by a factor of n, and this is good enough reason for preferring it.

The reader will have noted that the examples in this paper are both small and simple. They were chosen as such in order to simplify the exposition, and the various studies given have been selected in order to indicate the varieties of numerical behavior which can occur as a result of applying an acceleration technique, rather than to display the technique in a favorable light. Integro-differential equations have not been treated, but it is unlikely that they would have introduced new points. A more serious consequence of the simplification has been that only square and triangular two-dimensional arrays have been dealt with, whereas in practice

much more varied arrays are likely to occur. But the Samelson inverse can quite clearly be applied to such arrays so that no generality has been lost. This also applies to arrays of higher dimensions.

It is appropriate at this juncture to mention that the author's experience in a large number of practical cases has been that the accelerated results were never actually worse than the original iterated sequence, though whether the gain in accuracy is worth the extra computation involved depends upon the example being considered.

Many standard iterative processes of numerical analysis are of course already rapidly convergent. In particular, it seems (although the author is unaware of any general theorem to this effect) that if the partial differential equation

$$\alpha\{\phi(x, y), x, y\} = \beta\{\phi(x, y), x, y\} \tag{60}$$

is iterated according to the scheme

$$\alpha\{\phi_{r+1}(x, y), x, y\} = \beta\{\phi_r(x, y), x, y\} \tag{61}$$

and the equation

$$\alpha\{\phi(x, y), x, y\} = 0 \tag{62}$$

and (60) are of the same type (e.g., both elliptic) then the resulting scheme converges, and it is only when equations (60) and (62) do not conform that difficulty may arise. This is exemplified by Table 11 and may be illustrated further by means of the examples

$$\phi_{xx} + \phi_{yy} = 2\{\phi_x^2 - \phi_y^2\} \tag{63}$$

$$\phi_{xy} = \phi_x^2 - \phi_y^2 \tag{64}$$

which both have the same solution as (48), viz., $x/(x + y)$. This phenomenon also occurs in the method of Rayleigh and Jansen [21] for the iterative solution of the equation of potential flow.

Acceleration techniques may have interesting repercussions upon the practical treatment of partial differential equations, for suppose that (60) is a hyperbolic partial differential equation with curved characteristics, while (62) has linear characteristics. The evaluation of the right-hand side of (60) along linear characteristics will not, in general, present much difficulty and avoids the usual circumstantial interpolation. If equation (60) is nonlinear some sort of iteration is usual. The normal objection to such a procedure is that it does not converge, but this, as Table 11 for example indicates, is no longer a critical point.

However, it would be most desirable for further work to be done on the convergence behavior of iterative processes and acceleration techniques. The latter is, of course, an interesting, useful, and largely unexplored domain of research. It is also desirable that further work should be done on the exploitation of these techniques and this paper was written in the hope that the subject would be thrown open to a larger forum of experimentation.

6. Acknowledgement. The results displayed in Tables 2 to 4 were produced by the author using the XI computer in Amsterdam, and those displayed in Tables

5 to 12 were produced by the author with the aid of the ALGOL compiler constructed by E. W. Dijkstra and J. Zonneveld for this computer. The author is grateful to Prof. Dr. K. Samelson for describing his vector inverse to him.

Mathematisch Centrum
2e Boerhaavestraat 49,
Amsterdam-O.

1. P. WYNN, "On a device for computing the $e_m(S_n)$ transformation," MTAC, v. 10, 1956, p. 91.
2. P. WYNN, "The rational approximation of functions which are formally defined by a power series expansion," *Math. Comp.*, v. 14, 1960, p. 147.
3. D. SHANKS, "Non-linear transformations of divergent and slowly convergent sequences," *J. Math. Phys.*, v. 34, 1955, p. 1.
4. T. J. STIELTJES, "Sur la réduction en fraction continue d'une serie précédant suivant les puissances descendants d'une variable," *Ann. Fac. Sci. Univ. Toulouse*, v. 3, 1889, p. 1.
5. P. WYNN, "A comparison between the numerical performances of the Euler transformation and the Epsilon algorithms," *Chiffres*, v. 4, 1961, p. 23.
6. P. WYNN, "On repeated application of the ϵ-algorithm," *Chiffres*, v. 4, 1961, p. 19.
7. F. L. BAUER, *Connections between the q-d algorithm of Rutishauser and the ϵ-algorithm of Wynn*, A technical report prepared under the sponsorship of the Deutsche Forschungsgemeinschaft, Project No. BA:106, Nov. 1957.
8. J. R. SCHMIDT, "On the numerical solution of linear simultaneous equations by an iterative method," *Phil. Mag. Ser.* 7, v. 32, 1941, p. 362.
9. E. BODEWIG, "A practical refutation of the iteration method for the algebraic eigenproblem," MTAC, v. 8, 1945, p. 237.
10. J. MORRIS, "An escalator process for the solution of linear simultaneous equations," *Phil. Mag. Ser.* 7, v. 37, 1946, p. 106.
11. E. R. LOVE, "The electrostatic field of two equal circular coaxial conducting discs," *Quart. J. Mech. Appl. Math.*, v. 2, 1949, p. 428.
12. L. FOX & E. T. GOODWIN, "The numerical solution of non-singular integral equations," *Philos. Trans. Roy. Soc. London. Ser. A*, v. 245, 1953, p. 501–534.
13. A. C. AITKEN, "On Bernoulli's numerical solution of algebraic equations," *Proc. Roy. Soc. Edinburgh*, v. 46, 1926, p. 287.
14. F. G. FRIEDLANDER, "The reflection of sound pulses by convex parabolic reflectors," *Proc. Cambridge Philos. Soc.*, v. 37, 1941, p. 134.
15. C. W. CLENSHAW & F. W. J. OLVER, "Solution of differential equations by recurrence relations," *MTAC*, v. 5, 1951, p. 34.
16. L. FOX & E. T. GOODWIN, "Some new methods for the numerical integration of ordinary differential equations," *Proc. Cambridge Philos. Soc.*, v. 45, 1949, p. 373.
17. L. FOX & J. C. P. MILLER, "Table making for large arguments, the exponential integral," *MTAC*, v. 5, 1951, p. 163.
18. R. A. BUCKINGHAM, *Numerical Methods*, Pitman, London, 1957, p. 504–505.
19. D. R. HARTREE, *Numerical Analysis*, Oxford, 1952, p. 238.
20. P. WYNN, "Singular rules for certain non-linear algorithms," to appear.
21. R. SAUER, *Einführung in die theoretische Gasdynamik*, Springer, Berlin, 1951.
22. C. LANCZOS, "Linear systems in self-adjoint form," *Amer. Math. Monthly*, v. 65, 1958, p. 665.
23. P. WYNN, "On the solution of a certain boundary-value problem," *Nord. Tid. Inf. Behand.*, v. 2, 1962, p. 61.

Computing Error Bounds in Solving Linear Systems

By J. Schröder

1. Introduction. Let there be given a system of m linear algebraic equations for m unknowns

$$Gu = r \tag{1.1}$$

and consider an iteration procedure

$$u_{n+1} = Mu_n + s \qquad (n = 0, 1, 2, \cdots) \tag{1.2}$$

such that the system

$$u = Mu + s \tag{1.3}$$

is equivalent to (1.1). All elements of the m-dimensional vectors and the $m \times m$-matrices which occur are assumed to be real.

In the space of m-dimensional vectors $u = (u^i)$, $v, \cdots$ we define an order relation and an absolute value by writing

$$u \leqq v \quad \text{if and only if} \quad u^i \leqq v^i \qquad (i = 1, 2, \cdots, m),$$

and

$$|u| = (|u^i|).$$

Similarly, for $m \times m$-matrices $A = (a_{ij})$, $B, \cdots$ we use the notation

$$A \leqq B \quad \text{if and only if} \quad a_{ij} \leqq b_{ij} \quad (i, j = 1, 2, \cdots, m),$$

and

$$|A| = (|a_{ij}|).$$

Let B denote a matrix such that

$$|M| \leqq B$$

for M in equation (1.2) (use, for example, $B = |M|$ if M is explicitly known). If there exists a vector v_0 such that

$$|u_p - u_{p+1}| \leqq v_0 - Bv_0 \quad \text{for some index } p, \tag{1.4}$$

then the given equation has a solution u^* for which

$$|u^* - u_{p+n}| \leqq B^n v_0 \qquad (n = 0, 1, 2, \cdots) \tag{1.5}$$

holds (Theorem 1).

We present a method for computing a vector v_0 with the property (1.4). The calculation of v_0 and the vectors $B^n v_0$ in (1.5) constitutes an iterative procedure parallel to approximation procedure (1.2). This estimation procedure, described

Received August 2, 1961. This work was sponsored by the Mathematics Research Center, U.S. Army, Madison, Wisconsin.

at the end of Section 2, can be programmed for computers as easily as approximation procedure (1.2).

If the matrix B is irreducible and noncyclic and if the spectral radius $\rho(B)$ of the matrix B satisfies

$$\rho(B) < 1, \tag{1.6}$$

then for some p the method yields a vector v_0 which satisfies (1.4) (Theorem 3).

Section 4 is concerned with the methods of simultaneous and successive displacements (with B chosen as in (4.4)). Some numerical examples (up to 15 unknowns) are given. In these examples, the estimation procedure yields a suitable vector v_0, i.e., an error bound, after at most three steps, thus the time needed for estimation is considerably smaller than the time needed for solving the given system by iteration (one need not start the estimation procedure and the approximation procedure simultaneously).

For many of the known estimation methods one has to calculate an upper bound σ of $\rho(B)$ and this bound has to satisfy the condition

$$\sigma < 1. \tag{1.7}$$

For example, this inequality (1.7) may be the row-sum criterion for the method of successive displacements. In this paper condition (1.7) is weakened to (1.6) where $\rho(B)$ need not be known.

Of course, condition (1.6) still restricts the class of iteration procedures (1.2) to which the estimation method can be applied. Note that the procedure (1.2) converges for an arbitrary vector u_0 if and only if

$$\rho(M) < 1$$

where the spectral radius $\rho(M)$ of M satisfies

$$\rho(M) \leqq \rho(|M|) \leqq \rho(B).$$

However, every convergence condition which works with upper bounds b_{ij} of the moduli $|m_{ij}|$ instead of with elements m_{ij} cannot be weaker than condition (1.6).

For example, when solving difference equations for the Laplace equation by the method of successive displacements, one has, in general, $M \geqq O$ and $\rho(M) < 1$, thus, $\rho(B) < 1$ for $B = M$. In the case of the biharmonic equation, however, the method of successive displacements in general yields $\rho(|M|) > 1$.

2. Derivation of the Method. Let R be the set of m-dimensional real vectors $u = (u^i)$, $v, \cdots$, and let $A = (a_{ij})$, $B, M, \cdots$ denote real $m \times m$-matrices. The notation $u \leqq v$, $|u|$, $A \leqq B$, and $|A|$ shall be defined as in the introduction.

Consider an equation

$$u = Mu + s \tag{2.1}$$

where M denotes a given matrix, s is a given vector and u is unknown. With $Tu = Mu + s$ we can write this equation as $u = Tu$.

Let B denote a fixed matrix such that

$$|M| \leqq B, \tag{2.2}$$

then let $H[u, v]$ be the function

$$H[u, v] = \tfrac{1}{2}(B + M)\, u - \tfrac{1}{2}(B - M)\, v + s.$$

This function is increasing in u and decreasing in v, and for $u = v$ we get $H[u, u] = Tu$.

We consider the iteration procedure

$$x_{n+1} = H[x_n, y_n], \qquad y_{n+1} = H[y_n, x_n] \qquad (n = 0, 1, 2, \cdots). \tag{2.3}$$

Because of the described properties of $H[u, v]$, for this procedure the following statements hold (see [5]).

Let the conditions

$$x_0 \leqq x_1, \qquad x_0 \leqq y_0, \qquad y_1 \leqq y_0$$

be satisfied.

Then, the vectors x_n and y_n $(n = 0, 1, 2, \cdots)$ defined by (2.3) *satisfy the inequalities*

$$x_0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq y_n \leqq \cdots \leqq y_2 \leqq y_1 \leqq y_0. \tag{2.4}$$

Moreover, the sequences $\{x_n\}$ and $\{y_n\}$ converge to limit-vectors x^* and y^*, respectively, such that

$$x^* = H[x^*, y^*], \qquad y^* = H[y^*, x^*] \tag{2.5}$$

and

$$x_n \leqq x^* \leqq y^* \leqq y_n \qquad (n = 0, 1, 2, \cdots). \tag{2.6}$$

The inequalities (2.4) can be proved by induction, and the convergence of the sequences $\{x_n\}$ and $\{y_n\}$ then follows from the fact that these sequences are monotonic and bounded.

Adding the two equations in (2.5) and noting that (2.6) holds, we get, in addition, the following statement:

The vector

$$u^* = \tfrac{1}{2}(x^* + y^*)$$

is a solution of the given equation (2.1), *and for this solution the estimate*

$$x_n \leqq u^* \leqq y_n \qquad (n = 0, 1, 2, \cdots)$$

holds.

Now let $\{u_n\}$ and $\{v_n\}$ be sequences of vectors which satisfy the equations

$$u_{n+1} = Mu_n + s \qquad (n = 0, 1, 2, \cdots) \tag{2.7}$$

and

$$v_{n+1} = Bv_n \qquad (n = 0, 1, 2, \cdots), \tag{2.8}$$

respectively. Then, the vectors x_n and y_n defined by

$$u_n = \tfrac{1}{2}(x_n + y_n), \qquad v_n = \tfrac{1}{2}(y_n - x_n) \qquad (n = 0, 1, 2, \cdots),$$

i.e.,

$$x_n = u_n - v_n, \qquad y_n = u_n + v_n \qquad (n = 0, 1, 2, \cdots),$$

satisfy (2.3). We reformulate the main results stated above in terms of the vectors u_n and v_n instead of x_n and y_n.

THEOREM 1. *Let the conditions*

$$v_0 \geqq o \quad \text{and} \quad |u_0 - u_1| \leqq v_0 - v_1 \tag{2.9}$$

be satisfied.

Then, the vectors u_n and v_n $(n = 0, 1, 2, \cdots)$ defined by (2.7) *and* (2.8) *satisfy the inequalities*

$$v_n \geqq o, \qquad |u_n - u_{n+1}| \leqq v_n - v_{n+1} \qquad (n = 0, 1, 2, \cdots). \tag{2.10}$$

Moreover, the sequences $\{u_n\}$ and $\{v_n\}$ converge to vectors u^* and v^*, respectively, such that

$$u^* = Mu^* + s \quad \text{and} \quad v^* = Bv^*,$$

and for u^* the error estimate

$$|u^* - u_n| \leqq v_n \qquad (n = 0, 1, 2, \cdots) \tag{2.11}$$

holds.

From Theorem 1 we start to develop a method of error estimation for the iteration procedure (2.7). Clearly, we can replace the vectors u_i in the theorem, which occur by u_{i+p} (p denoting a fixed non-negative integer). The assumption (2.9) then takes the form

$$v_0 \geqq o, \quad |u_p - u_{p+1}| \leqq v_0 - v_1 \tag{2.12}$$

and the estimate (2.11) becomes

$$|u^* - u_{n+p}| \leqq v_n \qquad (n = 0, 1, 2, \cdots). \tag{2.13}$$

When the vectors u_p and u_{p+1} are known one could try to construct a suitable vector v_0 using the special properties of the given problem. However, the error estimate would then, in general, be much more complicated than the calculation of the approximations u_n because one can easily program the procedure (2.7) for computers. Therefore, we will establish a method of error estimation which also can be programmed quite easily.

A Method of Error Estimation for the Iteration Procedure $u_{n+1} = Mu_n + s$. Starting with some vector

$$w_0 \geqq o \quad (\text{for example } w_0 = o)$$

and using the formula

$$w_{n+1} = Bw_n + \delta_{n+1} \quad \text{with} \quad \delta_{n+1} = |u_{n+1} - u_n| \tag{2.14}$$

calculate vectors w_n up to the first index $n = p$ for which

$$w_p \geqq w_{p+1}, \tag{2.15}$$

provided that such an index exists. Calculate in addition vectors z_n $(n = p, p + 1, \cdots)$ defined by

$$z_p = w_p, \qquad z_{n+1} = Bz_n \qquad (n = p, p + 1, \cdots).$$

THEOREM 2. *If there exists an index p such that* (2.15) *holds, then the sequence* $\{u_n\}$ *converges to a solution* $u^* = Mu^* + s$ *and*

$$|u^* - u_n| \leqq z_n \qquad (n = p, p+1, \cdots). \tag{2.16}$$

Proof. Suppose p is an index such that (2.15) holds. Then, let $\{v_n\}$ denote the sequence defined by (2.8) for $v_0 = w_p$. Clearly, one has $v_0 \geqq o$. Moreover, also the second inequality in (2.12) is satisfied because this inequality is equivalent to the condition (2.15). Thus the sequence $\{u_n\}$ converges to a solution u^* which satisfies the inequality (2.13), and this inequality is equivalent to (2.16).

3. Theoretical Investigation of the Method. We now investigate the conditions under which the method of error estimation described above will be successful.

For this, we assume that the matrix B is irreducible (non-decomposable) and noncyclic. Then, also because $B \geqq O$, according to Frobenius [3] the following statements hold.

Matrix B has an eigenvalue $\lambda > 0$, called the maximal root of B, such that λ is greater than the modulus of each other eigenvalue of B. Corresponding to λ there exists an eigenvector $\varphi = (\varphi^i)$ with

$$\varphi^i > 0 \qquad (i = 1, 2, \cdots, m), \tag{3.1}$$

and there are no eigenvectors or principal vectors (generalized eigenvectors) corresponding to λ which are linearly independent of φ.

Thus, each vector $u \in R$ can be written as a sum

$$u = u^{(1)}\varphi + \psi \tag{3.2}$$

where $u^{(1)}$ is a constant and $\psi = \psi(u)$ is a linear combination of the eigenvectors and principal vectors of B belonging to eigenvalues different from λ.

Then let P_1 and P_2 denote the two projection matrices defined by

$$P_1 u = u^{(1)}\varphi \qquad (\text{for } u \in R), \qquad P_2 = I - P_1,$$

and moreover let

$$B_1 = BP_1, \qquad B_2 = BP_2.$$

Then the equation

$$B = B_1 + B_2$$

represents a spectral decomposition of matrix B with B_1 belonging to λ and B_2 belonging to the remainder of the spectrum of B. We have

$$B^n u = B_1{}^n u + B_2{}^n u = \lambda^n u^{(1)}\varphi + B_2{}^n u. \tag{3.3}$$

Let $\rho(A)$ denote the spectral radius of a matrix A, i.e., the maximum of the moduli of its eigenvalues. Then, we have $\rho(B_1) = \rho(B) = \lambda$ and

$$\rho(B_2) < \lambda. \tag{3.4}$$

If u is a vector such that $u \geqq o$, then $B^n u \geqq o$ $(n = 0, 1, 2, \cdots)$ and it follows from (3.3) that

$$u^{(1)}\varphi + (\lambda^{-1}B_2)^n u \geqq o. \tag{3.5}$$

Because of (3.4) the inequality (3.5) yields $u^{(1)}\varphi \geqq o$ for $n \to \infty$, hence $u^{(1)} \geqq 0$. This proves the statement that

$$u \geqq o \quad \textit{implies} \quad u^{(1)} \geqq 0. \tag{3.6}$$

If $u^i > 0$ for all $i = 1, 2, \cdots, n$, then there exists a constant $\alpha > 0$ such that $u \geqq \alpha\varphi$ and we now derive from (3.3)

$$u^{(1)}\varphi + (\lambda^{-1}B_2)^n u = \lambda^{-n}B^n u \geqq \lambda^{-n}\alpha B^n\varphi = \alpha\varphi.$$

Hence, for $n \to \infty$ we obtain the result

$$u^i > 0 \ (\textit{for all } i = 1, 2, \cdots, n) \ \textit{implies} \ u^{(1)} > 0. \tag{3.7}$$

We write

$$u > o \quad \text{if and only if} \quad u \geqq o \text{ and } u \neq o,$$

and we now also prove that

$$u > o \quad \textit{implies} \quad u^{(1)} > 0. \tag{3.8}$$

Let $\mu > \lambda$, then the equation

$$\sum_{i=0}^{\infty} \mu^{-i}B^i = (I - \mu^{-1}B)^{-1}$$

holds, and the matrix $(I - \mu^{-1}B)^{-1}$ has all elements positive because B is irreducible. (This follows from Theorem 2.2 in [7], applied to $A = \mu I$, $z = \varphi$ and B as defined above.) Therefore, if $u > o$, all components of the vector

$$v = (I - \mu^{-1}B)^{-1}u = \sum_{i=0}^{\infty} \mu^{-i}B^i u$$

are positive, hence it follows from (3.7) that $v^{(1)} > 0$. Finally, we conclude from the equation

$$v^{(1)}\varphi = P_1 v = P_1 \sum_{i=0}^{\infty} \mu^{-i}B^i u = \sum_{i=0}^{\infty} \mu^{-i}\lambda^i u^{(1)}\varphi = u^{(1)}(1 - \lambda\mu^{-1})^{-1}\varphi$$

and that $u^{(1)} > 0$.

THEOREM 3. *Suppose that matrix B is irreducible and noncyclic, and assume moreover that the maximal root $\lambda = \rho(B)$ satisfies*

$$\lambda < 1. \tag{3.9}$$

Then there exists an index p such that (2.15) holds.

Proof. Consider the sequence $\{\omega_n\}$ defined by

$$\omega_0 = w_0, \qquad \omega_{n+1} = B\omega_n + \delta_{n+1} \qquad (n = 0, 1, 2, \cdots).$$

We shall prove that

$$\omega_n - \omega_{n+1} \geqq o \quad \text{for } n \text{ large enough}, \tag{3.10}$$

which is sufficient for the existence of an index p with the desired property.

We first suppose that $\omega_0 = w_0 = o$. Then, if $u_0 = u_1$ the inequality (2.15) is satisfied for $p = 0$. Therefore, we will assume that $\delta_1 = |u_1 - u_0| > o$.

The difference $\omega_n - \omega_{n+1}$ can be written as

$$\omega_n - \omega_{n+1} = (I - B)(B^{n-1}\delta_1 + B^{n-2}\delta_2 + \cdots + \delta_n) - \delta_{n+1} \quad (n = 1, 2, \cdots)$$

with

$$\delta_j = |M^{j-1}\epsilon_1| \qquad (j = 1, 2, \cdots) \tag{3.11}$$

and

$$\epsilon_1 = u_1 - u_0 .$$

In the following we consider two cases described by $\rho(M) < \lambda$ and $\rho(M) = \lambda$, respectively.

Case I. Let

$$\rho(M) < \lambda. \tag{3.12}$$

Using (3.3), we decompose $\omega_n - \omega_{n+1}$ into two summands

$$\omega_n - \omega_{n+1} = S_n^{\,1} + S_n^{\,2} \qquad (n = 1, 2, \cdots) \tag{3.13}$$

with

$$S_n^{\,1} = (1-\lambda)(\lambda^{n-1}\delta_1^{(1)} + \lambda^{n-2}\delta_2^{(1)} + \cdots + \delta_n^{(1)})\varphi,$$

$$S_n^{\,2} = (P_2 - B_2)\eta_n - \delta_{n+1}$$

and

$$\eta_n = B_2^{\,n-1}\delta_1 + B_2^{\,n-2}\delta_2 + \cdots + \delta_n .$$

Because of (3.6), all coefficients $\delta_j^{(1)}$ are non-negative. Therefore, we get

$$\lambda^{-n+1}S_n^{\,1} \geqq (1 - \lambda)\delta_1^{(1)}\varphi \qquad (n = 1, 2, \cdots), \tag{3.14}$$

and from (3.1), (3.8), and (3.9) we show that the vector on the right side of (3.14) has all components positive.

Because the spectral radii of B_2 and M are smaller than λ and because δ_j is of the form (3.11), the series $\sum_{n=0}^{\infty}\lambda^{-n}B_2^{\,n}$ and

$$\sum_{n=1}^{\infty}\lambda^{-n+1}\delta_n \tag{3.15}$$

converge. Therefore, the product series

$$\sum_{n=1}^{\infty}\lambda^{-n+1}\eta_n = \sum_{i=0}^{\infty}\lambda^{-i}B_2^{\,i}\cdot\sum_{j=1}^{\infty}\lambda^{-j+1}\delta_j \tag{3.16}$$

also converges. In particular, the summands $\lambda^{-n+1}\delta_n$ and $\lambda^{-n+1}\eta_n$ of the series (3.15) and (3.16) converge to the null vector. Thus, we have

$$\lim_{n\to\infty}\lambda^{-n+1}S_n^{\,2} = o, \tag{3.17}$$

and this relation, together with the inequality (3.14), indicates that (3.10) holds in Case I.

Case II. Suppose now that $\rho(M) = \lambda$. Then, according to a result of Wielandt [8], M can be written as

$$M = e^{i\alpha}D^{-1}BD \tag{3.18}$$

where α denotes a real number and D is a diagonal matrix with diagonal elements of modulus 1. Because M is supposed to be real the relation (3.18) holds for $\alpha = 0$ or $\alpha = \pi$ and a diagonal matrix D with diagonal elements ± 1.

In this case the vectors δ_j take the form

$$\delta_j = |M^{j-1}\epsilon_1| = |D^{-1}B^{j-1}D\,\epsilon_1| = |B^{j-1}\zeta| = |\lambda^{j-1}\zeta^{(1)}\varphi + B_2^{j-1}\zeta|$$

with $\zeta = D\,\epsilon_1$ $(j = 1, 2, \cdots)$.

If $\zeta^{(1)} = 0$, then we start again with the decomposition (3.13), for which (3.14) holds. As in Case I, we can prove again that (3.17) holds. For this, we now use $\delta_j = |B_2^{j-1}\zeta|$ and (3.4) instead of (3.11) and (3.12). As in Case I, (3.14) and (3.17) together yield (3.10).

Now let $\zeta^{(1)} \neq 0$. Then we have

$$\delta_j = |\zeta^{(1)}|\,\lambda^{j-1}\,|\psi_j| \quad \text{with} \quad \psi_j = \varphi + [\zeta^{(1)}]^{-1}(\lambda^{-1}B_2)^{j-1}\zeta \quad (j = 1, 2, \cdots).$$

Because of (3.1) and since the second summands of the ψ_j converge to o for $j \to \infty$ there exists a number j_0 such that $\psi_j \geqq o$ for $j \geqq j_0$; hence

$$\delta_j = |\zeta^{(1)}|\,\lambda^{j-1}\varphi + (sgn\ \zeta^{(1)})B_2^{j-1}\zeta \quad \text{for} \quad j \geqq j_0\,. \tag{3.19}$$

Suppose now that $n > j_0 + 1$. Then we write

$$\omega_n - \omega_{n+1} \quad \text{as a sum} \quad \omega_n - \omega_{n+1} = S_n^{\,3} + S_n^{\,4} \qquad (n > j_0 + 1)$$

where

$$S_n^{\,3} = (I - B)(B^{n-1}\delta_1 + B^{n-2}\delta_2 + \cdots + B^{n-j_0}\delta_{j_0})$$

and

$$S_n^{\,4} = (I - B)(B^{n-j_0-1}\delta_{j_0+1} + \cdots + \delta_n) - \delta_{n+1}\,.$$

Using (3.3) we decompose $S_n^{\,3}$ further into a sum

$$S_n^{\,3} = S_n^{\,31} + S_n^{\,32}$$

with

$$S_n^{\,31} = (1 - \lambda)(\lambda^{n-1}\delta_1^{(1)} + \lambda^{n-2}\delta_2^{(1)} + \cdots + \lambda^{n-j_0}\delta_{j_0}^{(1)})\varphi,$$

$$S_n^{\,32} = (I - B_2)B_2^{\,n-j_0}(B_2^{\,j_0-1}\delta_1 + \cdots + \delta_{j_0}).$$

Since the coefficients $\delta_j^{(1)}$ are non-negative, we get

$$\lambda^{-n+1}S_n^{\,31} \geqq (1 - \lambda)\delta_1^{(1)}\varphi \qquad (n > j_0 + 1), \tag{3.20}$$

where the vector on the right side of the inequality has all components positive.

The second summands $S_n^{\,32}$ satisfy

$$\lim_{n\to\infty} \lambda^{-n+1}S_n^{\,32} = o. \tag{3.21}$$

This follows from (3.4) because j_0 is a fixed number.

Using (3.19), we also split up $S_n^{\,4}$ into the following sum:

$$S_n^{\,4} = S_n^{\,41} + S_n^{\,42}$$

with

$$S_n^{41} = |\zeta^{(1)}|[(I-B)(\lambda^{j_0}B^{n-j_0-1} + \cdots + \lambda^{n-1}I)\varphi - \lambda^n\varphi]$$
$$= |\zeta^{(1)}|\lambda^{n-1}[(n-j_0)(1-\lambda) - \lambda]\varphi$$

and

$$S_n^{42} = (sgn\ \zeta^{(1)})[(I-B)(B^{n-j_0-1}B_2^{j_0} + \cdots + B_2^{n-1})\zeta - B_2^n\zeta]$$
$$= (sgn\ \zeta^{(1)})B_2^{n-1}[(n-j_0)(I-B_2) - B_4]\zeta.$$

We have

$$\lambda^{-n+1}S_n^{41} \geqq o \text{ for } n \text{ large enough,} \tag{3.22}$$

and from (3.4) we deduce that

$$\lim_{n\to\infty} \lambda^{-n+1}S_n^{42} = o. \tag{3.23}$$

Altogether, from the relations (3.20), (3.21), (3.22), and (3.23) it follows that (3.10) holds in Case II also.

Finally, let $\omega_0 = w_0 > o$. Then, the difference $\omega_n - \omega_{n+1}$ gets an additional summand

$$\tilde{S}_n = (I-B)B^n w_0 = (1-\lambda)\lambda^n w_0^{(1)}\varphi + (I-B_2)B_2^n w_0 \quad (n = 1, 2, \cdots).$$

These summands $\tilde{S}_n$ satisfy

$$\lim \lambda^{-n+1}\tilde{S}_n = (1-\lambda)\lambda w_0^{(1)}\varphi$$

where the vector on the right side has all components positive. Thus, (3.10) also holds in this case $w_0 > o$ (even if $u_1 - u_0 = o$, i.e., $\delta_1^{(1)} = 0$ in (3.14) and (3.20)).

COROLLARY. *Suppose that matrix B is irreducible and let $u_1 \neq u_0$. Then, condition* (3.9) *is necessary for the existence of a vector v_0 which satisfies*

$$v_0 \geqq o \quad \textit{and} \quad |u_0 - u_1| \leqq v_0 - Bv_0. \tag{3.24}$$

Remark. This corollary, together with Theorem 3, says, roughly speaking, that if B is irreducible and noncyclic, the method of estimation in Section 2 is always successful if one can get an estimate with Theorem 2.

Proof of the corollary. Since $u_1 \neq u_0$, it follows from (3.24) that

$$v_0 > o \quad \text{and} \quad (I-B)v_0 > o,$$

hence, in view of (3.8)

$$v_0^{(1)} > o \quad \text{and} \quad [(I-B)v_0]^{(1)} = (1-\lambda)v_0^{(1)} > o.$$

These two inequalities can hold only if $\lambda < 1$.

4. Applications to Numerical Examples. A given system of m linear equations

$$Gu = r \quad \text{with} \quad g_{ii} > 0 \qquad (i = 1, 2, \cdots, m) \tag{4.1}$$

can be written in the form (2.1) as follows. Let $G = D - C_1 - C_2$ where D is the diagonal matrix with diagonal elements g_{ii} and C_1 is some lower triangular matrix.

Then equation (4.1) is equivalent to (2.1) with $s = (D - C_1)^{-1}r$ and

$$M = (D - C_1)^{-1}C_2 . \tag{4.2}$$

In case $C_1 = O$ the iteration procedure

$$u_{n+1} = Mu_n + s \qquad (n = 0, 1, 2, \cdots) \tag{4.3}$$

is the method of simultaneous displacements of the system (4.1). On the other hand, if C_2 is an upper triangular matrix (4.3) represents the method of successive displacements for equation (4.1).

For matrix M given in equation (4.2) inequality (2.2) is satisfied for

$$B = (D - |C_1|)^{-1}|C_2| . \tag{4.4}$$

Clearly, one need not start the estimation procedure described at the end of Section 2, and the approximation procedure (4.3) simultaneously. One can define $w_0 = w_1 = \cdots = w_q$ (q denoting some non-negative integer), then start to calculate further vectors by equation (2.14). In this case, the estimation method is described by the following formulas for matrix B in equation (4.4):

$$w_q \geqq o \text{ (with } q \text{ a given non-negative integer)};$$

$$(D - |C_1|)\tau_{n+1} = |C_2|\, w_n\,, \quad w_{n+1} = \tau_{n+1} + \delta_{n+1} \quad (n = q, q+1, \cdots, p)$$

where $\delta_{n+1} = |u_{n+1} - u_n|$ and p is the smallest index such that $w_p \geqq w_{p+1}$;

$$z_p = w_p\,, \qquad (D - |C_1|)z_{n+1} = |C_2|\, z_n \qquad (n = p, p+1, \cdots).$$

This procedure has been programmed for the CDC 1604 Computer. Of course the index p and the bounds z_n depend on the value of the chosen q. To indicate this we write $p = p(q)$ and $z_n = z_n(q)$. According to Theorem 2 we have

$$|u^* - u_n| \leqq z_n(q) \quad \text{for} \quad n \geqq p(q).$$

The following examples have been calculated using the program mentioned above with $w_q = o$.

Example 1. The system $Gu = r$ with G and r as given in Table 1 consists of difference equations which approximate a certain boundary value problem for the Laplace equation [4]. We solve this system by the method of successive displacements. In the present case, we have $C_1 = |C_1|$ and $C_2 = |C_2|$, thus

$$M = B. \tag{4.5}$$

The starting vector u_0 as given in Table 2 is the (rounded) solution of difference equations for a smaller mesh width. In the same table are listed the exact errors $\zeta_n = u_n - u^*$ for some indices n and the corresponding bounds $z_n(q)$ for $q = 0$, 10, and 15. The indices $p(q)$ belonging to these values of q are

$$p(0) = 3, \qquad p(10) = 11, \qquad p(15) = 16.$$

This indicates that the estimation method was successful for $q = 0$ after three steps and for $q = 10$ and $q = 15$ after one step. Table 2 shows that the bounds for $q = 10$ and $q = 15$ are equal (up to nine decimals) but sharper than the bounds for $q = 0$. We learn from this that the estimation procedure should not be started

immediately ($q = 0$), but rather after the iteration process (4.3) has become "steady." That the bounds for $q = 10$ and $q = 15$ are almost equal is a consequence of (4.5) (in this connection, see also the next example). It would have been sufficient to start the estimation procedure at $q = 15$.

Example 2. Matrix G and vector r of the second example are given in Table 3. The corresponding equation $Gu = r$ consists of difference equations approximating a boundary value problem for the differential equation $\Delta\Delta u = \varphi(x, y)$ [1]. In general, condition $\rho(B) < 1$ is not satisfied for such problems. However, the estimation method works in this case with a few unknowns. We calculate this example in order to test how the bounds $z_n(q)$ behave if $\rho(M) < \rho(B) < 1$

Numerical results for $q = 0$, 10, and 25 are given in Table 4. The corresponding indices p are

$$p(0) = 2, \qquad p(10) = 12, \qquad p(25) = 27.$$

In this example, the largest index q gives the sharpest bounds. This can always be expected if $\rho(M) < \rho(B)$, because in this case the vectors $|\zeta_n|$ decrease faster than the bounds z_n. For example, if the equation $\det(M - \kappa I) = 0$ has a simple root κ_1 such that $|\kappa_1| = \rho(M)$ and all other eigenvalues of M have moduli smaller than $|\kappa_1|$, then we have in general

$$|\zeta_{n+1}| \approx \rho(M)\,|\zeta_n| \text{ for all sufficiently large } n \tag{4.6}$$

while

$$z_{n+1} \approx \rho(B) z_n \text{ for all sufficiently large } n.$$

The number of steps $p(q) - q$ from the beginning of the estimation procedure until success is achieved is the same for the two larger values of q. This phenomenon also occurred in Example 1. It can be explained by the fact that, in general, $\delta_{n+1} \approx \rho(M)\delta_n$ for n large enough.

Further Examples. In Example 3 we have solved a system of 15 difference equations which approximate the same boundary value problem as the equations in Example 1 and which are of the same type as those equations; however, in Example 3 the row-sum criterion is not satisfied. Compared with Example 1 there has been no essential difference in the behavior of the estimation procedure. Therefore, we give only a few numerical results in Table 5. Furthermore, we applied our method to a system $Gu = r$ with a 2-cyclic matrix G (Example 4) using the

TABLE 1
Coefficients of Example 1

Matrix G								Vector r
12	−1	−1	0	−2	−2	−2	−2	1
−2	12	0	0	0	−4	−4	0	1
−2	0	12	−2	−4	0	0	−2	0
0	0	−2	14	−2	0	0	0	0
−2	0	−2	−1	13	−1	0	−1	0
−2	−2	0	0	−1	12	−1	0	0
−2	−2	0	0	0	−1	12	−1	6
−4	0	−2	0	−2	0	−2	12	2

TABLE 2

Results of Example 1 (u_0 = starting vector, $\zeta_n = u_n - u^*$ = exact error, $z_n(q)$ = bounds for ζ_n)

i	1	2	3	4	5	6	7	8
$u_0{}^i$	0.399 6	0.465 5	0.215 5	0.053 9	0.153 0	0.215 5	0.715 5	0.502 2
$\zeta_3{}^i$	0.002 736 691	0.002 203 276	0.002 229 846	0.000 710 506	0.001 296 122	0.001 153 026	0.001 221 478	0.001 713 471
$z_3{}^i(0)$	0.009 759 418	0.008 644 262	0.007 523 341	0.002 570 770	0.004 525 089	0.004 433 653	0.004 477 596	0.006 008 977
ζ_{11}^i	0.000 005 720	0.000 004 516	0.000 004 751	0.000 001 519	0.000 002 721	0.000 002 393	0.000 002 550	0.000 003 577
$z_{11}^i(10)$	0.000 006 650	0.000 005 249	0.000 005 523	0.000 001 766	0.000 003 163	0.000 002 781	0.000 002 965	0.000 004 159
$z_{11}^i(0)$	0.000 020 738	0.000 016 370	0.000 017 223	0.000 005 508	0.000 009 863	0.000 008 673	0.000 009 244	0.000 012 968
ζ_{16}^i	0.000 000 121	0.000 000 096	0.000 000 101	0.000 000 032	0.000 000 058	0.000 000 051	0.000 000 054	0.000 000 075
$z_{16}^i(15)$	0.000 000 141	0.000 000 112	0.000 000 117	0.000 000 038	0.000 000 067	0.000 000 059	0.000 000 063	0.000 000 088
$z_{16}^i(10)$	0.000 000 141	0.000 000 112	0.000 000 117	0.000 000 038	0.000 000 067	0.000 000 059	0.000 000 063	0.000 000 088
$z_{16}^i(0)$	0.000 000 439	0.000 000 347	0.000 000 365	0.000 000 116	0.000 000 209	0.000 000 184	0.000 000 196	0.000 000 275

TABLE 3
Coefficients of Example 2

Matrix G				Vector r
12	−3	−3	1	1
−3	10	−2	−3	1
−3	−2	10	−3	1
2	−6	−6	11	1

TABLE 4
Results of Example 2 (u_o = starting vector, $\zeta_n = u_n - u^*$ = exact error, $z_n(q)$ = bounds for ζ_n)

i	1	2	3	4
u_0^i	1	1	1	1
ζ_2^i	0.079 260 342	0.097 059 129	0.175 309 129	0.117 789 897
$z_2^i(0)$	0.275	0.327 954 545	0.273 545 455	0.334 426 997
ζ_{12}^i	0.000 474 178	0.000 677 978	0.000 601 444	0.000 611 652
$z_{12}^i(10)$	0.000 861 331	0.001 076 452	0.000 967 820	0.001 140 014
$z_{12}^i(0)$	0.009 386 534	0.011 627 371	0.010 937 345	0.014 014 669
ζ_{27}^i	0.000 000 096	0.000 000 137	0.000 000 122	0.000 000 123
$z_{27}^i(25)$	0.000 000 174	0.000 000 217	0.000 000 196	0.000 000 230
$z_{27}^i(10)$	0.000 006 455	0.000 007 996	0.000 007 522	0.000 009 638
$z_{27}^i(0)$	0.000 076 075	0.000 094 237	0.000 088 644	0.000 113 585
ζ_{30}^i	0.000 000 018	0.000 000 025	0.000 000 022	0.000 000 023
$z_{30}^i(25)$	0.000 000 062	0.000 000 076	0.000 000 072	0.000 000 092
$z_{30}^i(10)$	0.000 002 464	0.000 003 053	0.000 002 871	0.000 003 679
$z_{30}^i(0)$	0.000 029 040	0.000 035 973	0.000 033 838	0.000 043 358

procedure of simultaneous displacements. In this case matrix B in (4.4) was 2-cyclic also, and the estimation procedure was unsuccessful.

General Remarks Concerning Practical Application. The sharpness of the bounds $z_n(q)$ depends on the chosen index q. In general, for larger q one gets sharper bounds after the estimation procedure has been successful, i.e., for $n \geqq p(q)$. Only if $\rho(M) = \rho(B)$ can one expect that the bounds are almost equal for different q, provided q is so large that the iteration process has become "steady." Moreover, the time needed for the estimation becomes smaller for larger q. Therefore, the best way might be to start the estimation procedure with an index q such that δ_q is smaller than a suitably chosen bound. Then, the approximations are improved still more for $n > q$.

There are several further possible ways to verify the program. For example, one may use the fact that in general the differences $p(q) - q$ become equal for q large enough. One may start at some index q_1 in order to find $p(q_1) - q_1$, then stop the estimation procedure and start it again with a suitable index q_2 such that $p(q_2) \approx q_2 + (p(q_1) - q_1)$.

In our examples we used the starting vector $w_q = o$. In other cases however,

TABLE 5
Some Results of Example 3

u_0^1	$= 0.39256$
u_{18}^1	$= 0.384\ 711\ 925$
$z_{18}^1(15)$	$= 0.000\ 007\ 026$
u_{28}^1	$= 0.384\ 707\ 090$
$z_{28}^1(25)$	$= 0.000\ 000\ 145$
$z_{28}^1(15)$	$= 0.000\ 000\ 145$
u_{30}^1	$= 0.384\ 707\ 035$
$z_{30}^1(25)$	$= 0.000\ 000\ 067$
$z_{30}^1(15)$	$= 0.000\ 000\ 067$

for a vector

$$w_q = \beta\delta_q$$

with $\beta > 0$ the corresponding number p might be much smaller than for $w_q = o$.

For example, let

$$M \geqq O \quad \text{and} \quad B = M.$$

Then, for q large enough the difference $u_q - u_{q+1}$ in general is approximately proportional to φ. Suppose that $u_q - u_{q+1} = \varphi$ and choose $w_q = \beta\delta_q = \beta\varphi$. Then the first index $n = p$ for which $w_p \geqq w_{p+1}$ holds is the smallest integer such that

$$p \geqq \lambda[(1-\lambda)^{-1} - \beta] + q \quad \text{and} \quad p \geqq q.$$

For $\beta = 0$ and λ very close to 1 this is a large number, for $\beta > (1-\lambda)^{-1}$ however, one has $p = q$. Of course, for β much larger than $(1-\lambda)^{-1}$ the bound $w_p = w_q$ is not sharp.

For example, solving large systems of difference equations for the Laplace equation by the method of successive displacements one may choose

$$w_q = (1-\kappa)^{-1}\delta_q$$

where κ approximates the corresponding eigenvalue λ. Such an approximation κ is known in many cases.

Round-off errors have not been considered in our program. However, we believe that it is not difficult to do this if suitable subroutines are available. It certainly is not necessary to take into consideration all round-off errors which occur in the entire approximation and estimation procedure. One has to do this only for the last step.

Let n_0 be the index up to which the approximations and bounds have been calculated. Then, consider the vectors u_{n_0-1} and z_{n_0-1} as they are computed with a certain number of digits. If one can show that

$$z_{n_0-1} \geqq o \quad \text{and} \quad |u_{n_0-1} - u_{n_0}| \leqq z_{n_0-1} - z_{n_0} \tag{4.7}$$

hold, then as a consequence of Theorem 1, applied to u_{n_0-1} and z_{n_0-1} instead of u_0

and v_0, there exists a solution u^* such that $|u^* - u_{n_0}| \leqq z_{n_0}$ and $|u^* - u_{n_0-1}| \leqq z_{n_0-1}$.

If the entire approximation and estimation process could be done without round-off errors, then, certainly, inequalities (4.7) would be satisfied. This follows from the statement (2.10) in Theorem 1. Therefore, in general, one can expect that (4.7) can be proved also for the vectors u_{n_0-1} and z_{n_0-1} which are actually computed.

In (4.7), u_{n_0} and z_{n_0} do not denote the vectors numerically calculated, but the exact vectors defined by

$$u_{n_0} = Mu_{n_0-1} + s, \qquad z_{n_0} = Bz_{n_0-1}. \tag{4.8}$$

Thus, one has to estimate the round-off errors which occur in computing u_{n_0} and z_{n_0} by (4.8).

In order to do this, one may, for example, reckon with pairs of numbers (instead of numbers) in (4.8). For example, a subroutine for calculation with pairs of numbers has been written for the IBM 650 Computer and this subroutine has been successfully applied for error estimation for certain differential equations [6]. A similar subroutine has been developed by G. E. Collins for the IBM 704 Computer (1959). The method of calculating with pairs of numbers is called "interval arithmetic" by Collins. This method has been used for desk computers by Dwyer [2] under the name "range arithmetic."

Institut für Angewandte Mathematik
Universität Hamburg

Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

1. L. Collatz, *The Numerical Treatment of Differential Equations*, third edition, Springer-Verlag, Berlin, 1961, p. 391.
2. P. S. Dwyer, *Linear Computations*, John Wiley & Sons, Inc., New York, 1951.
3. G. Frobenius, "Über Matrizen aus nichtnegativen Elementen," *S.-B. Preuss. Akad. Wiss. Berlin*, 1912, p. 456–477.
4. J. Schröder, "Zur Lösung von Potentialaufgaben mit dem Differenzenverfahren," *Z. Angew. Math. Mech.*, v. 34, 1954, p. 241–253.
5. J. Schröder, "Anwendung von Fixpunktsätzen bei der numerischen Behandlung nichtlinearer Gleichungen in halbgeordneten Räumen," *Arch. Rational Mech. Anal.*, v. 4, 1959, p. 177–192.
6. J. Schröder, "Fehlerabschätzung mit Rechenanlagen bei gewöhnlichen Differentialgleichungen erster Ordnung," *Numer. Math.*, v. 3, 1961, p. 39–61, 125–130.
7. J. Schröder, "Lineare Operatoren mit positiver Inversen," *Arch. Rational Mech. Anal.*, v. 8, 1961, p. 408–434.
8. H. Wielandt, "Unzerlegbare nichtnegative Matrizen," *Math. Z.*, v. 52, 1950, p. 642–648.

A Note on Finite Difference Methods for Solving the Eigenvalue Problems of Second-Order Differential Equations

By M. R. Osborne

1. Introduction. In many cases of importance a finite difference approximation to the eigenvalue problem of a second-order differential equation reduces the problem to that of solving the eigenvalue problem of a tridiagonal matrix having the Sturm property. In this paper we illustrate the use of the Rayleigh quotient for obtaining a quadratically convergent iteration to the eigenvalues of such a matrix.

If the finite difference approximation is to give a tridiagonal matrix having the Sturm property then there must be restrictions on the coefficients in the differential equation and on the form of the boundary conditions. The type of problem we consider is defined in Section 2.

The Rayleigh quotient iteration is derived in Section 3 where we also relate it to a class of quadratically convergent iterations. This section is included at the suggestion of a referee to whom I am indebted for several other constructive suggestions and, in particular, for bringing to my attention the paper by J. W. Cooley [1].

The Rayleigh quotient iteration is exemplified in Section 4. We show that it can be conveniently linked to a difference correction procedure for estimating the error in the solution of the finite difference approximation. The stability of the iteration is discussed in Section 5; and a routine which we have used on an electronic computer is described in Section 6.

2. Finite Difference Approximations. In the usual notation the standard method of approximating to a second-order differential equation using finite difference formulas on a grid of equispaced points equates $h^2 \frac{d^2}{dx^2}$ with δ^2, and $h \frac{d}{dx}$ with $\mu\delta$ where h is the grid spacing. For example, the equation

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + (\lambda q(x) + r(x))y = 0 \tag{2.1}$$

becomes

$$(\delta^2 + hp(x)\mu\delta + h^2(\lambda q(x) + r(x))y = 0, \tag{2.2}$$

and when written out at the jth grid point this gives

$$\left(1 - \frac{h}{2}p_j\right)y_{j-1} - (2 - h^2(\lambda q_j + r_j))y_j + \left(1 + \frac{h}{2}p_j\right)y_{j+1} = 0. \tag{2.3}$$

We assume boundary conditions

$$Ay(0) + B\frac{dy(0)}{dx} = 0, \qquad \text{and} \quad y(e) = 0. \tag{2.4}$$

Received February 7, 1961.

Setting $h = e/n$, $x_j = jh$ we have the finite difference approximations to (2.4) in the form

$$y_{-1} = y_1 + \alpha y_0, \qquad \alpha = 2h\frac{A}{B}, \tag{2.5}$$
$$y_n = 0.$$

Combining the boundary conditions (2.5) with the set of equations (2.3) for $0 \leqq j \leqq n-1$ we obtain an algebraic eigenvalue problem which we write as

$$(A + h^2\lambda Q)\mathbf{y} = 0. \tag{2.6}$$

Here $\mathbf{y}$ is the vector with components y_j, $0 \leqq j \leqq n-1$, Q is a diagonal matrix whose $(j+1)$st element is q_j, and A is a tridiagonal matrix. We adopt the notation that the nonzero elements in the $(j+1)$st row of A are c_j, a_j, b_j.

We assume that $q(x)$ is strictly positive in $0 \leqq x \leqq e$, and that h can be chosen sufficiently small so that $|\frac{h}{2}p_j| < 1$, $1 \leqq j \leqq n-1$. From these assumptions it follows that the principal minors of the determinant of $A + h^2\lambda Q$ have the Sturm property [2, p. 181–2], so that a technique of counting signs and bisection can be used to calculate the eigenvalues of (2.6).

In general the matrix A is not symmetric. It can, however, be made symmetric by premultiplying it by a certain diagonal matrix D. If the $(j+1)$th element of D is d_j then the elements of the $(j+1)$th row of DA are d_jc_j, d_ja_j, and d_jb_j so that DA is symmetric if $d_jc_j = d_{j-1}b_{j-1}$. This gives a recurrence for d_j when one value which can be freely chosen is fixed. Thus we can always write equation (2.6) in the form

$$(DA + h^2\lambda DQ)\mathbf{y} = 0 \tag{2.7}$$

where DA is symmetric, and where DQ is diagonal with positive elements.

When $p(x)$ is identically zero a more accurate finite difference formula (the Numerov formula) can be used to approximate to equation (2.1). This gives the difference equations

$$\left(1 + \frac{h^2}{12}(\lambda q_{i-1} + r_{i-1})\right)y_{i-1} - \left(2 - \frac{5h^2}{6}\lambda q_i + r_i\right)y_i + \left(1 + \frac{h^2}{12}(\lambda q_{i+1} + r_{i+1})\right)y_{i+1} = 0. \tag{2.8}$$

In this case the appropriate boundary condition at $x = 0$ is

$$y_{-1}\left(1 + \frac{h^2}{6}(\lambda q_{-1} + r_{-1})\right) = y_1\left(1 + \frac{h^2}{6}(\lambda q_1 + r_1)\right) + \alpha y_0. \tag{2.9}$$

These equations combine to give the eigenvalue problem

$$(M + h^2\lambda N)\,\mathbf{y} = 0 \tag{2.10}$$

where both M and N are tridiagonal matrices. It is not possible in general to make both M and N symmetric by premultiplying by a diagonal matrix.

3. Initial Value Techniques for Correcting an Approximate Eigenvalue. The methods we consider in this section have as a common feature that they seek to correct λ by using an approximate eigenvector $\mathbf{V}$ which is obtained as the solution to an initial value problem. To compute $\mathbf{V}$ we choose V_0 arbitrarily, then

$$b_0 V_1 = -(h^2\lambda q_0 + a_0) V_0 ,$$

and

$$b_i V_{i+1} = -(h^2\lambda q_i + a_i) V_i - c_i V_{i-1} , \qquad i = 2, 3, \cdots, n-2. \tag{3.1}$$

It is important to note that by suitable choice of V_0 we can arrange for any one of the components of $\mathbf{V}$ to be fixed independent of the value of λ.

From equation (3.1) we see that the components of $A\mathbf{V}$ are

$$b_i V_{i+1} + a_i V_i + c_i V_{i-1} = -h^2\lambda q_i V_i , \qquad i = 0, 1, \cdots, n-2,$$

$$c_{n-1} V_{n-2} + a_{n-1} V_{n-1} = -h^2\lambda q_{n-1} V_{n-1} + (h^2\lambda q_{n-1} + a_{n-1}) V_{n-1} + c_{n-1} V_{n-2} ,$$

so that $\mathbf{V}$ also satisfies

$$(A + h^2\lambda Q)\mathbf{V} = \beta \mathbf{e}_n \tag{3.2}$$

where $\mathbf{e}_n$ is a vector with 1 in the nth place, and zero in all other positions. Also β vanishes whenever λ is an eigenvalue of (2.6) so that $\beta(\lambda)$ is proportional to the characteristic polynomial of (2.6).

One method of improving λ is to calculate the Rayleigh quotient of equation (2.7) using the approximate eigenvector $\mathbf{V}$. In this case the Rayleigh quotient is

$$\frac{\mathbf{V}'DA\mathbf{V}}{\mathbf{V}'DQ\mathbf{V}} = \frac{\mathbf{V}^{*\prime}A\mathbf{V}}{\mathbf{V}^{*\prime}Q\mathbf{V}} \quad \text{where} \quad \mathbf{V}^* = D\mathbf{V}.$$

Making use of equation (3.2) we have

$$\frac{\mathbf{V}^{*\prime}A\mathbf{V}}{\mathbf{V}^{*\prime}Q\mathbf{V}} = -h^2\lambda + \frac{\beta V^*_{n-1}}{\mathbf{V}^{*\prime}Q\mathbf{V}}. \tag{3.3}$$

It will be seen that (3.3) gives the correction to λ in a very economical form.

Another possibility is to apply Newton's method to find the zeros of $\beta(\lambda)$. This suggests that $\lambda - \beta/\frac{d\beta}{d\lambda}$ is a better approximation to the desired eigenvalue. We can calculate $\frac{d\beta}{d\lambda}$ by differentiating equation (3.2). This gives us

$$(A + h^2\lambda Q)\frac{d\mathbf{V}}{d\lambda} + h^2 Q\mathbf{V} = \frac{d\beta}{d\lambda}\mathbf{e}_n . \tag{3.4}$$

If we take the scalar product of equation (3.4) with $\mathbf{e}_n$ we have

$$c_{n-1}\frac{dV_{n-2}}{d\lambda} + (a_{n-1} + h^2\lambda q_{n-1})\frac{dV_{n-1}}{d\lambda} + h^2 q_{n-1} V_{n-1} = \frac{d\beta}{d\lambda}. \tag{3.5}$$

This equation forms the basis of the procedure described by Fox (see [3], Chapter 8, and references given there). He calculates $\frac{d\mathbf{V}}{d\lambda}$ by choosing the scale of $\mathbf{V}$ so that V_0 in equation (3.1) is independent of λ. This gives $\frac{dV_0}{d\lambda} = 0$, and we have

$$b_0 \frac{dV_1}{d\lambda} = -(h^2\lambda q_0 + a_0)\frac{dV_0}{d\lambda} - h^2 q_0 V_0 = -h^2 q_0 V_0,$$

$$b_i \frac{dV_{i+1}}{d\lambda} = -(h^2\lambda q_i + a_i)\frac{dV_i}{d\lambda} - c_i \frac{dV_{i-1}}{d\lambda} - h^2 q_i V_i, \tag{3.6}$$

$$i = 1, 2, \cdots, n-2.$$

However, if we take the scalar product of equation (3.4) with $\mathbf{V}^*$ we find, on noting that $\mathbf{V}^*$ satisfies

$$(A' + h^2\lambda Q)\mathbf{V}^* = \beta D\mathbf{e}_n = d_{n-1}\beta\mathbf{e}_n,$$

that

$$d_{n-1}\beta\mathbf{e}_n' \frac{d\mathbf{V}}{d\lambda} + h^2\mathbf{V}^{*\prime}Q\mathbf{V} = V_{n-1}^* \frac{d\beta}{d\lambda}. \tag{3.7}$$

The correction formula given by Cooley [1] can be derived from (3.7) by choosing the scale of $\mathbf{V}$ so that V_{n-1} is independent of λ. Then $\frac{dV_{n-1}}{d\lambda} = \mathbf{e}_n' \frac{d\mathbf{V}}{d\lambda} = 0$, and $\frac{d\beta}{d\lambda} = h^2 \frac{\mathbf{V}^{*\prime}Q\mathbf{V}}{\mathbf{V}_{n-1}^*}$. The correction to λ is $-h^{-2}\frac{V_{n-1}^*\beta}{\mathbf{V}^{*\prime}Q\mathbf{V}}$. It is identical with that given by the Rayleigh quotient (3.3).

It does not follow from this result that Fox's procedure is equivalent to an application of the Rayleigh quotient because the value of β is dependent on the way in which the scale is fixed in each method (although the correction to λ in each case is independent of the choice of scale), and the scale factors are in general different functions of λ. We make this point clear by relating the principal minors of the determinant of (2.6) to the successive terms in the recurrence (3.1).

The principal minors of (2.6) satisfy the recurrence

$$\begin{aligned} D_0 &= 1. \\ D_1 &= (a_0 + h^2\lambda q_0)D_0, \\ D_{i+1} &= (a_i + h^2\lambda q_i)D_i - c_i b_{i-1} D_{i-1}, \\ D &= \det(A + h^2\lambda Q) = (a_{n-1} + h^2\lambda q_{n-1})D_{n-1} - c_{n-1}b_{n-2}D_{n-2}. \end{aligned}$$

If we put $V_0 = sD_0$ where s is the scale factor, then

$$b_0 V_1 = -sD_1,$$

and

$$\begin{aligned} b_1 V_2 &= -(h^2\lambda q_1 + a_1)V_1 - c_1 V_0 \\ &= (h^2\lambda q_1 + a_1)\frac{sD_1}{b_0} - c_1 sD_0 = \frac{sD_2}{b_0}, \end{aligned}$$

so that $b_0 b_1 V_2 = sD_2$. This procedure can be continued, giving

$$sD_i = (-1)^i b_0 b_1 \cdots b_{i-1} V_i, \tag{3.8}$$

and, in particular,

$$D = \det(A + h^2\lambda Q) = (-1)^{n-1}\frac{b_0 b_1 \cdots b_{n-2}}{s}\beta. \tag{3.9}$$

If V_0 is fixed independent of λ then s is a constant, and β differs from the characteristic polynomial only by a constant multiplier. Therefore, Fox's method is equivalent to applying Newton's method to the characteristic polynomial. If V_{n-1} is fixed independent of λ then, by (3.8), s is inversely proportional to D_{n-1} so that β differs from D/D_{n-1} only by a constant multiplier. Now, by a well-known result for Sturm sequences, the zeros of D_{n-1} interlace those of D so that for the Rayleigh quotient to give a useful correction to λ it is necessary that λ lie inside the interval containing the desired eigenvalue bounded by the pair of adjacent zeros of D_{n-1}. No such restriction applies to Fox's iteration.

The results obtained in this section can readily be extended to the Numerov difference approximation (equation (2.10)). The initial value procedure gives a vector $\mathbf{V}$ satisfying

$$(M + h^2\lambda N)\mathbf{V} = \beta \mathbf{e}_n , \tag{3.10}$$

and the only change is in the definition of $\mathbf{V}^*$ which we take as the solution of

$$(M' + h^2\lambda N')\mathbf{V}^* = \beta^* \mathbf{e}_n . \tag{3.11}$$

We can calculate $\mathbf{V}^*$ by the initial value procedure. The value of β^* then depends on the choice of scale. The Rayleigh quotient is given by

$$\frac{\mathbf{V}^{*\prime} M \mathbf{V}}{\mathbf{V}^{*\prime} N \mathbf{V}} = -h^2\lambda + \frac{\beta V_n^*}{\mathbf{V}^{*\prime} N \mathbf{V}} , \tag{3.12}$$

and this determines the correction to λ. The application of Newton's method proceeds exactly as before.

4. The Rayleigh Quotient Exemplified. In this section we exemplify the use of the Rayleigh quotient by applying it to calculate the fundamental eigenvalue of the standard finite difference approximation (equation (2.3)) to the differential equation

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + \lambda y = 0 \tag{4.1}$$

with the boundary conditions $\dfrac{dy(0)}{dx} = 0$, $y(1) = 0$. As $\dfrac{dy}{dx} = 0$ when $x = 0$ we see that the required solution to (4.1) satisfies

$$2\frac{d^2y}{dx^2} + \lambda y = 0 \tag{4.2}$$

at this point. Also it is an even function about $x = 0$ so that the finite difference approximation (equation (2.5)) to the boundary condition $\dfrac{dy(0)}{dx} = 0$ is exact.

In this case Q is the unit matrix. Taking $n = 4$, $h = \frac{1}{4}$ we find that A is the matrix

$$\begin{bmatrix} -4 & 4 & & \\ \frac{1}{2} & -2 & \frac{3}{2} & \\ & \frac{3}{4} & -2 & \frac{5}{4} \\ & & \frac{5}{6} & -2 \end{bmatrix}.$$

It is readily verified that a suitable diagonal matrix such that DA is symmetric is

$$D = \begin{bmatrix} \frac{1}{2} & & & \\ & 4 & & \\ & & 8 & \\ & & & 12 \end{bmatrix}.$$

In the table that follows we give the result of the computation. We use equation (3.1) to run forward the approximate eigenvectors (rounding to six decimals). The corrections to λ are shown in the final column.

$h^2\lambda$	$y(0)$	$y(\frac{1}{4})$	$y(\frac{1}{2})$	$y(\frac{3}{4})$	β	$\dfrac{\beta V_n^*}{\mathbf{V}^{*\prime}\mathbf{V}}$
.312500	1.000000	.921875	.703776	.396973	−.083412	−.040742
.353242	1.000000	.911689	.667555	.332428	.008867	.004058
.349184	1.000000	.912704	.671138	.338718	.000120	.000055
.349129	1.000000	.912718	.671187	.338804	.000001	—

The quadratic nature of the convergence is clearly indicated.

For this example the Rayleigh quotient requires about half as much arithmetic as Fox's method. It has another advantage for the vector $\mathbf{V}^*$ is readily to hand if it is required to compute an estimate of the deviation of the computed λ from the exact eigenvalue of the differential equation by using the difference correction (see reference [3]). Returning to our example we set

$$\mathfrak{C}y_i = \frac{1}{12}\left\{\frac{2}{i}\mu\delta^3 + \delta^4\right\} y_i$$

except when $i = 0$ when

$$\mathfrak{C}y_0 = \tfrac{1}{6}\delta^4 y_0 .$$

As the finite difference boundary conditions are exact in this case an estimate of the correction to λ is given by

$$h^2 \Delta\lambda = \frac{\mathbf{V}^{*\prime}\mathfrak{C}\mathbf{y}}{\mathbf{V}^{*\prime}\mathbf{y}} . \tag{4.3}$$

The difference table is

i	y_i	δ	δ^2	δ^3	$\mu\delta^3$	δ^4
−1	.912718					
		−.087282			0	
0	1.000000		−.174564			.040630
		−.154249		.020315		
1	.912718		−.241531		.041856	.043082
		−.332383		.063399		
2	.671187		−.090852		.073914	.021034
		−.338804		.084431		
3	.338804		−.006421		.083071	−.002720
		−.263514		.081711		
4	0		.075290			
		−.263514				
5	−.263514					

Equation (4.3) gives us $h^2\Delta\lambda = .01184$. The basic difference equation gives $\lambda = 5.59$. Taking into account the difference correction gives $\lambda = 5.776$ which compares much more favourably with the exact value ($\lambda = 5.781 \cdots$).

5. The Problem of Stability. The crucial factor to be considered in suggesting the Rayleigh quotient procedure as a general routine is the stability of the calculation of the approximate eigenvector; it is well known that we cannot use the procedure summarized in equation (3.1) in general as the resulting vector may well differ substantially from the exact eigenvector. There is, however, numerical evidence that Fox's iteration works (and gives quadratic convergence) for the eigenvalue at least in some cases of instability in the calculation of the vector (see [4]). It is doubtful if the Rayleigh quotient iteration can be successful in these circumstances, and it is interesting that the numerical evidence would seem to indicate that there is a root of D_{n-1} very close to the desired root of D in the direction of increasing λ. In many cases this evidence can be supported by heuristic argument.

In solving problems associated with second-order differential equations one often knows beforehand that instability is likely for part of the range for which the solution is defined. For example, the desired solution may have the character of an exponential decay in some region where the second solution to the differential equation is changing rapidly, and we may anticipate that the solutions of our second-order difference equation mirror this behavior. However, if instability is likely when running forward due to a rapidly increasing spurious solution then we may expect stability if we reverse the direction of the computation. This device works well within the limits of its applicability.

With little loss of convenience we may march the approximate eigenvector back from the last row of the matrix as well as forward from the first row. If we match the computations by making their rth components agree then we have solved

$$(A + h^2\lambda Q)\mathbf{V} = \beta \mathbf{e}_r \tag{5.1}$$

where $1 \leqq r \leqq n$. The analysis of Section 3 can be applied to (5.1). We find that Newton's method applied to $\beta(\lambda)$ with v_0 fixed independent of λ (Fox's correction) is equivalent to applying Newton's method to D/E_{n-r+1}, while the Rayleigh quotient is equivalent to Newton's method applied to $D/D_{r-1}E_{n-r+1}$. Here E_{n-r+1} is the principal minor with diagonal elements $(a_r + h^2\lambda q_r), \cdots, (a_{n-1} + h^2\lambda q_{n-1})$.

If it is necessary to do more than march from both ends to achieve stability then the initial value techniques lose much of their simplicity. In this case we are forced to solve an equation of the form

$$(A + h^2\lambda Q)\mathbf{V} = \mathbf{k} \tag{5.2}$$

where we have considerable freedom in the choice of $\mathbf{k}$. The calculation of an eigenvector of $A + h^2\lambda Q$ from (5.2) has been discussed in detail by Wilkinson in reference [5]. In particular, he suggests that $\mathbf{k}$ at any stage should be the approximate eigenvector calculated in the previous iteration.

In considering finite difference approximations to eigenvalues of higher-order differential equations we find that A is no longer tridiagonal, and that the Rayleigh quotient procedure cannot be readily combined with an initial value procedure for the desired eigenvector. Fox's procedure can be modified to apply in this case [3] and [4]. The cost is the addition of further variational equations similar to (3.6). However, the Rayleigh quotient is defined whenever the approximate vectors

$\mathbf{V}$ and $\mathbf{V}^*$ are available, and Wilkinson's experience in [5] suggests the iteration

$$(A + h^2\lambda_{s-1}Q)\mathbf{V}_s = \mathbf{V}_{s-1},$$

$$(A' + h^2\lambda_{s-1}Q)\mathbf{V}_s^* = \mathbf{V}_{s-1}^*, \tag{5.3}$$

$$h^2\lambda_s = h^2\lambda_{s-1} - \frac{\mathbf{V}_s^{*\prime}\mathbf{V}_{s-1}}{\mathbf{V}_s^{*\prime}Q\mathbf{V}_s}.$$

Setting $A + h^2\lambda Q = LU$, then

$$\mathbf{V}_s = U^{-1}L^{-1}\mathbf{V}_{s-1},$$

and

$$\mathbf{V}_s^* = (L^{-1})'(U^{-1})'\mathbf{V}_{s-1}^* \tag{5.4}$$

so that each iteration (5.3) requires only one matrix factorization. For numerical stability it would be necessary to use the method of triangular factorization with interchanges given by Wilkinson in [6, p. 18–19].

6. A Computational Procedure. This paper is based on experience gained over several years in solving differential equation eigenvalue problems which arise in the theory of the propagation of sound in the ocean treated as a layered fluid medium (the theoretical background to this work has been described in [7]). Here one boundary is taken at an infinite depth (which is treated as very great in the numerical work); the differential equations have, in general, mixed spectra. Our concern has been to calculate the eigenvalues and eigenfunctions in the discrete spectrum.

The differential equation has the form

$$\frac{d^2y}{dx^2} + (\lambda + q(x))y = 0$$

so that the Numerov difference approximation (equation (2.10)) is applicable, and we solve this in the following stages.

(I) We begin by using the standard finite difference approximation. We use the Sturm count procedure to roughly isolate an eigenvalue; and we then use the Rayleigh quotient formula to obtain a quadratically convergent iteration to it. The use of the standard equation instead of the Numerov equation at this stage approximately halves our arithmetic.

(II) We calculate an $O(h^2)$ improvement to this value of λ using the difference correction technique exactly as described in Section 4. This gives an eigenvalue having the same kind of order of magnitude deviation from the exact value as does the corresponding eigenvalue of (2.10).

(III) We use the Rayleigh quotient iteration (3.12) to calculate the accurate eigenvalue of equation (2.10) from the value of λ obtained in stage (II). We also obtain a useful improvement in the accuracy of the eigenvector. If further refinement is required we can calculate an improvement to λ (and, if desired, to $\mathbf{y}$) by a further stage of difference correction.

Using values of n up to 100 we have been able to obtain stable computations by marching forward the calculation of the eigenvector from each boundary point, and matching at a suitably chosen interior point. There is physical evidence to

support the use of these tactics because regions of low acoustic pressure can occur near the surface of the ocean as well as at great depth. In no case have more than two iterations been required in stage (III) of the computation. Comparison of the values of λ obtained at the various stages of the computation gives us useful information about the accuracy of our approximations.

University of Reading
England

1. J. W. Cooley, "Improved formula for solving central fields," *Math. Comp.*, v. 15, 1961, p. 363–374.
2. A. S. Householder, *Principles of Numerical Analysis*, McGraw-Hill, 1953.
3. L. Fox, *The Numerical Solution of Two-point Boundary Value Problems*, Oxford, 1957.
4. L. Fox, "Some numerical experiments with eigenvalue problems in ordinary differential equations," included in *Boundary Problems in Differential Equations* (Editor R. E. Langer), University of Wisconsin Press, 1960.
5. J. H. Wilkinson, "Calculation of eigenvectors of codiagonal matrices," *Comput. J.*, v. 1, 1958, p. 90–96.
6. National Physical Laboratory, *Modern Computing Methods, Notes on Applied Science No.* 16, Her Majesty's Stationery Office, London, 1961.
7. M. R. Osborne, "The propagation of sound in a layered fluid medium," *Quart. J. Mech. Appl. Math.*, v. 13, n. 4, 1960, p. 472–86.

A Method for Computing the Circular Coverage Function

By A. R. DiDonato and M. P. Jarnagin

1. Introduction. In this paper an efficient method is described for the numerical evaluation, with a high-speed digital computer, of a special case of the integral of an uncorrelated bivariate Gaussian distribution centered at the origin over the area of an arbitrarily placed circle in the plane. This function, popularly known as the circular coverage function or as the non-central chi-square distribution for two degrees of freedom*, can be written as

$$P(R, D) \equiv \frac{1}{2\pi\sigma_x\sigma_y}\iint_S \exp\left\{-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2+\left(\frac{y}{\sigma_y}\right)^2\right]\right\}dxdy, \tag{1}$$

where S is the circle: $(x-h)^2+(y-k)^2=(\sigma R)^2$, where $\sigma_x=\sigma_y=\sigma$, and σD is the radial distance from the origin to the center (h, k) of the circle of integration, S. Because of the equivalence mentioned above, a great deal of published literature applies. The papers [13], [15], suggested by the referee, list a large number of such references.

The average computing time for the calculation of the integral in equation (1) to six decimal digits, by the method of this paper, is six milliseconds on the IBM 7090 and ten milliseconds on NORC. An extensive inverse table, which is described in the last section of this paper and which is given in [4], has been computed with R as a function of P and D. A condensed version, Table 1, is presented herein.

In the general case [3], [11] suppose the uncorrelated bivariate Gaussian distribution centered at the origin of an Oxy Cartesian coordinate system has standard deviations σ_x, σ_y along the x and y axes respectively, and that the integral of this function is to be evaluated over a circle of radius $\bar{R}$ with center at (h, k). Then the probability, P, can be written in polar coordinates accordingly:

$$P\left(\frac{\bar{R}}{\sigma_x}, \frac{\bar{R}}{\sigma_y}, \frac{h}{\sigma_x}, \frac{k}{\sigma_y}\right) = \frac{1}{2\pi\sigma_x\sigma_y}\int_0^{\bar{R}}\int_0^{2\pi} \cdot \exp\left\{-\frac{1}{2}\left[\left(\frac{h+r\cos\theta}{\sigma_x}\right)^2+\left(\frac{k+r\sin\theta}{\sigma_y}\right)^2\right]\right\} rdrd\theta, \tag{2}$$

where $x-h=r\cos\theta$, $y-k=r\sin\theta$, $0\leqq r\leqq\bar{R}$, $0\leqq\theta\leqq 2\pi$.

If

$$h=k=0,$$

a special case identified as the $V(K, c)$ or elliptical normal probability function (sometimes known by other titles, for example, the generalized circular error function) [4], [5], [6], [10], [14], [15], [16], [18] follows, i.e.,

$$P\left(\frac{\bar{R}}{\sigma_x}, \frac{\bar{R}}{\sigma_y}, 0, 0\right) \equiv V(K, c) = \frac{1}{c}\int_0^K \exp\left(-\frac{B}{2}r^2\right) I_0\left(\frac{Ar^2}{2}\right) rdr, \tag{3}$$

Received July 27, 1961.

* The equivalence between the function $P(R, D)$ of equation (1) and the non-central chi-square distribution is evident from equation (2) in [13].

where

$$0 \leqq c \equiv \frac{\sigma_y}{\sigma_x} \leqq 1, \qquad K \equiv \bar{R}/\sigma_x, \qquad A \equiv \frac{1-c^2}{2c^2}, \qquad B \equiv \frac{1+c^2}{2c^2}.$$

$I_0(x)$ is the modified Bessel function of the first kind of order zero, [8]. Equation (3) is derived by setting $h = k = 0$ in equation (2), by using the trigonometric identity $1(\pm) \cos 2\theta = 2\begin{pmatrix}\cos^2\theta \\ \sin^2\theta\end{pmatrix}$, and by introducing an integral expression for $I_0(x)$ which is given by

$$I_0(x) = \frac{1}{\pi}\int_0^{\pi} \exp(-x\cos\theta)\, d\theta. \tag{4}$$

Equation (4) can be derived from Example 1 (ii), page 62, in [8].

If

$$\sigma_x = \sigma_y = \sigma,$$

in equation (2), the distribution is circular normal. In this case, in which h and k are arbitrary, the center of the circle of integration can always be taken as offset a distance of σD from the origin along the positive x axis by simply introducing a rotation of axes through the angle arc $\tan\left(\frac{k}{h}\right)$. Moreover, by introducing the integral expression for $I_0(x)$ as given by equation (4), the circular coverage function, $P(R, D)$, [1], [4], [6], [7], [9], [12], [13], [14], [17], is obtained from equation (2), i.e.,

$$P\left(\frac{\bar{R}}{\sigma_x}, \frac{\bar{R}}{\sigma_x}, \frac{h}{\sigma_x}, \frac{k}{\sigma_x}\right) \equiv P(R, D) = \exp(-D^2/2)\int_0^R \exp(-r^2/2) I_0(rD) r\, dr, \tag{5}$$

where $R \equiv \bar{R}/\sigma_x$, $D^2 \equiv (h^2 + k^2)/\sigma_x^2$.

The function $\partial P(R, D)/\partial R$ is required for computing the inverse function, $R(P, D)$, by the Newton-Raphson procedure (Appendix C, [4]) and is also of use in computing $P(R, D)$ itself (see equation (9)). This function is obtained straightforwardly from equation (5) as

$$\frac{\partial P}{\partial R} = R\exp\left(-\frac{R^2 + D^2}{2}\right) I_0(RD). \tag{6}$$

It is apparent by comparing equations (6), (9) that $\partial P/\partial R$ can be computed simultaneously with $P(R, D)$.

In a previous paper, [18], a very efficient computing method was described for calculation of the $V(K, c)$ function. The success of the method warranted consideration of extending the technique to the $P(R, D)$ function. This is not as straightforward as for $V(K, c)$; nevertheless, it is easily possible because of the existence of a simple functional relationship, equation (9), between $P(R, D)$ and $V(K, c)$.

2. The Relationship between P(R, D) and V(K, c). The relationship between P and V can be derived by utilizing two preliminary results which are given by Fettis, in terms of $q \equiv 1 - P$, in equations (I-35) and (I-44) in [6]. They can be

stated in terms of P as:

$$P(R,D) - P(D,R) = \pm V\left(|R-D|, \frac{|R-D|}{R+D}\right) \quad \begin{matrix}(+) \text{ if } R > D \\ (-) \text{ if } R < D,\end{matrix} \tag{7}$$

$$P(R,D) + P(D,R) = 1 - \exp\left(-\frac{R^2+D^2}{2}\right) I_0(RD). \tag{8}$$

Equation (8) is easily derived. The origin of equation (7) is not known to the authors. The referee has pointed out that a geometrical proof was given by Dr. David C. Kleinecke of the University of California in 1955. (See also paper I of [15], page 613). Mr. Fettis has kindly placed at the disposal of the authors some correspondence which indicates that the relationship was given in a Sandia Corporation working paper in 1952, and that it was believed to have been originally derived in a British publication by using power series.

It follows by adding the corresponding sides of equations (7) and (8) that*

$$P(R,D) = \frac{1}{2}\left[1 - \exp\left(-\frac{R^2+D^2}{2}\right) I_0(RD) \pm V\left(|R-D|, \frac{|R-D|}{R+D}\right)\right] \tag{9}$$

$$\begin{matrix}(+) \text{ if } R > D \\ (-) \text{ if } R < D.\end{matrix}$$

Thus, the $P(R, D)$ function is computable at virtually the same speed as $V(K, c)$, since the second term in the brackets turns out to be a by-product of the recurrence relations which are used to compute V in the last term. Consequently, if there exists a satisfactory computing program for the V function, a computing program of equal merit can be realized for the $P(R, D)$ function.

3. Recurrence Relations. The V function that appears as the last term of equation (9) is identified with equation (3) by setting

$$K = |R - D|, \qquad c = |R - D|/(R + D).$$

It follows that

$$A = \frac{2RD}{(R-D)^2}, \qquad B = \frac{R^2+D^2}{(R-D)^2},$$

where it is assumed $R \neq D$. If $R = D$, then, from equation (7), $V\left(|R-D|, \frac{|R-D|}{R+D}\right)$ vanishes and $P(R, D)$ is given by the first two terms of equation (9).

The two series representations for $V\left(|R-D|, \frac{|R-D|}{R+D}\right)$ from which the basic recurrence relations are derived are given by:

* Guenther recently (see equation (2) in [9]) derived an equation for $P(R, D)$ in terms of $I_0(x)$ and the incomplete gamma function, which can be shown to be equivalent to equation (9) of the present paper. However, he did not exploit his relationship from the point of view of developing an efficient program for a high-speed digital computer.

$$V\left(|R-D|, \frac{|R-D|}{R+D}\right) = \frac{|R^2-D^2|}{RD} \sum_{n=0}^{\infty} \left(\frac{1}{n!}\right)^2 \cdot \int_0^{RD/2} \exp\left(-\frac{R^2+D^2}{RD} w\right) w^{2n}\, dw \equiv \sum_{n=0}^{\infty} T_{2n}, \tag{10}$$

$$V\left(|R-D|, \frac{|R-D|}{R+D}\right) = 1 - \frac{|R^2-D^2|}{4RD\sqrt{\pi}} \sum_{n=0}^{N} \frac{[(2n)!]^2}{2^{4n}(n!)^2} \cdot \int_{2RD}^{\infty} \exp\left[-\frac{(R-D)^2}{4RD} w\right] w^{-\left(\frac{2n+1}{2}\right)}\, dw = 1 - \sum_{n=0}^{N} M_{2n+1}. \tag{11}$$

The detailed derivations of equations (10), (11) are given in [4]. Briefly, to obtain equation (10), introduce a variable of integration transformation

$$w = Ar^2/4 \tag{12}$$

into the integral of equation (3), then replace $I_0(2w)$ by its Taylor series expansion (see page 14, [8]),

$$I_0(2w) = \sum_{n=0}^{\infty} \left(\frac{1}{n!}\right)^2 \left(\frac{2w}{2}\right)^{2n}, \tag{13}$$

which is convergent for all values of w, and subsequently reverse the order of integration and summation, which can be justified by application of the Weierstrass "M" test. In order to derive equation (11) introduce a variable of integration transformation

$$w = Ar^2 \tag{14}$$

into the integral of equation (3) and use the fact that

$$\frac{1}{2Ac} \int_0^{\infty} \exp\left(-\frac{Bw}{2A}\right) I_0\left(\frac{w}{2}\right) dw = 1, \tag{15}$$

(See page 76, [8]). In the resulting integral expression, call it J, with upper and lower limits of integration of infinity and AK^2 respectively, replace $I_0\left(\frac{w}{2}\right)$ by its asymptotic expansion (see page 58, [8]), i.e.,

$$I_0\left(\frac{w}{2}\right) \approx \frac{\exp(w/2)}{\sqrt{2\pi(w/2)}} \sum_{n=0}^{N} \frac{[(2n)!]^2}{2^{4n}(n!)^2} (2w/2)^{-n}, \tag{16}$$

which is valid for sufficiently large w and finite N; subsequently interchange the order of integration and summation. The interchange is justified for all values of $(2RD)$ for which equation (16) is valid because of the existence of the integral J (see page 17, [2]).

The substitution of equations (13), (16) into equation (6) gives analogous series representations for $\partial P/\partial R$, i.e.,

$$\frac{\partial P}{\partial R} = R \sum_{n=0}^{\infty} \bar{S}_{2n}, \tag{17}$$

$$\frac{\partial P}{\partial R} \approx R\left(\frac{1}{4RD} \sum_{n=0}^{N} \bar{X}_{2n+1}\right), \tag{18}$$

where

$$\bar{S}_{2n} \equiv \exp\left(-\frac{R^2+D^2}{2}\right)\left(\frac{1}{n!}\right)^2\left(\frac{RD}{2}\right)^{2n}, \qquad n \geqq 0, \tag{19}$$

$$\bar{X}_{2n+1} \equiv \frac{2}{\sqrt{\pi}} \exp\left[-\frac{(R-D)^2}{2}\right] \frac{[(2n)!]^2}{2^{4n}(n!)^3} (2RD)^{-\left(\frac{2n-1}{2}\right)}, \qquad n \geqq 0, \tag{20}$$

following the notation of [4], in which there are slight distinctions between $\bar{S}_{2n}$, $\bar{X}_{2n+1}$, $\bar{Y}_{2n-1}$, and the corresponding unbarred variables used with $V(K, c)$ and $\partial V/\partial K$.

Thus two schemes are used to compute P. If

$$2RD \leqq M \quad (M \text{ is a positive constant}), \tag{21}$$

then with reference to equations (10) and (17)

$$T_{2n} = \left(\frac{2n-1}{2n}\right)\left(\frac{2RD}{R^2+D^2}\right)^2 T_{2n-2} - \frac{|R^2-D^2|}{R^2+D^2}\left(1+\frac{4n}{R^2+D^2}\right)\bar{S}_{2n}, \qquad n \geqq 1, \tag{22}$$

$$\bar{S}_{2n} = \left(\frac{RD}{2n}\right)^2 \bar{S}_{2n-2}, \qquad n \geqq 1, \tag{23}$$

where the necessary initial terms are given by

$$T_0 = \frac{|R^2-D^2|}{R^2+D^2}\left[1-\exp\left(-\frac{R^2+D^2}{2}\right)\right] = \frac{|R^2-D^2|}{R^2+D^2}(1-\bar{S}_0), \tag{24}$$

$$\bar{S}_0 = \exp\left(-\frac{R^2+D^2}{2}\right). \tag{25}$$

The following brief comments are made on the derivation of recurrence relations (22) and (23). Fuller details are given in [4]. From equations (13) and (19), $\bar{S}_{2n}$ is the general term in the series obtained by multiplying every term of the Taylor series for $I_0(RD)$ by $\exp[-(R^2+D^2)/2]$, and equations (23) and (25) are obtained immediately. If T_{2n} is regarded as defined by equation (10), two successive integrations by parts give T_{2n} in terms of T_{2n-2}, R, D, and n, after which the term not containing T_{2n-2} can be written more concisely in terms of $\bar{S}_{2n}$, and equation (22) is the result.

These basic recurrence relations are cycled until

$$T_{2n} < \epsilon, \qquad \bar{S}_{2n} < \epsilon, \qquad (\epsilon > 0). \tag{26}$$

Then P and $\partial P/\partial R$ are given correctly to at least $(|\log_{10}\epsilon| - 1)$ decimal digits by

$$P(R, D) \approx \frac{1}{2}\left[1 - \sum_{n=0}^{N'} \bar{S}_{2n} \pm \sum_{n=0}^{N'} T_{2n}\right] \qquad \begin{matrix}(+) \text{ if } R > D \\ (-) \text{ if } R < D,\end{matrix} \tag{27}$$

$$\frac{\partial P}{\partial R} \approx R \sum_{n=0}^{N'} \bar{S}_{2n}. \tag{28}$$

If it is assumed that

$$2RD > M, \tag{29}$$

then with reference to equations (11) and (18)

$$M_{2n+1} = \frac{|R^2 - D^2|}{4RD}\,\bar{Y}_{2n-1} - \frac{(R-D)^2}{4RD}\left(\frac{2n-1}{2n}\right)M_{2n-1}, \qquad n \geqq 1, \tag{30}$$

$$\bar{Y}_{2n-1} = \frac{1}{4RD}\left(\frac{2n-1}{2n}\right)\bar{X}_{2n-1}, \qquad n \geqq 1, \tag{31}$$

$$\bar{X}_{2n+1} = (2n-1)\,\bar{Y}_{2n-1}, \qquad n \geqq 1, \tag{32}$$

where the initial terms are given by

$$\begin{aligned} M_1 &= \frac{1}{\sqrt{2RD}}\left(\frac{R+D}{\sqrt{2}}\right)\frac{2}{\sqrt{\pi}}\int_{\frac{|R-D|}{\sqrt{2}}}^{\infty}\exp(-y^2)\,dy \\ &= \frac{1}{\sqrt{2RD}}\left(\frac{R+D}{\sqrt{2}}\right)\left[1 - \operatorname{Erf}\left(\frac{|R-D|}{\sqrt{2}}\right)\right], \end{aligned} \tag{33}$$

$$\bar{X}_1 = \sqrt{2RD}\,\frac{2}{\sqrt{\pi}}\exp\left[-\frac{(R-D)^2}{2}\right]. \tag{34}$$

The following brief comments are made on the derivation of equations (30) to (33). Fuller details are given in [4]. From equations (16) and (20), $\bar{X}_{2n+1}$ is the general term in the expansion obtained by multiplying every term of the asymptotic expansion of $I_0(RD)$ by $4RD\exp[-(R^2+D^2)/2]$. Equations (31) and (32), which together form a recurrence relation generating $\bar{X}_{2n+1}$, are obtained immediately, the introduction of the variable $\bar{Y}_{2n-1}$ leading to a computationally efficient algorithm for the simultaneous evaluation of the last two terms in equation (9). If M_{2n+1} is regarded as defined by equation (11), an integration by parts gives M_{2n+1} in terms of M_{2n-1}, R, D, and n, after which the term not containing M_{2n-1} can be written more concisely in terms of $\bar{Y}_{2n-1}$, and recurrence relation (30) is the result. M_1, originally obtained by putting $n = 0$ in the definition of M_{2n+1}, is expressed in equation (33) in terms of the error function (see [3], equations (6)) by a transformation in which y is $(\frac{1}{2})|R - D|\sqrt{w/(RD)}$.

These basic recurrence relations are cycled until

$$M_{2n+1} < \epsilon, \qquad \bar{X}_{2n+1} < \epsilon, \qquad (\epsilon > 0). \tag{35}$$

Then $P(R, D)$ and $\partial P/\partial R$ are given correctly to $(|\log_{10}\epsilon| - 1)$ decimal digits by

$$P(R,D) \approx \frac{1}{2}\left[1 - \frac{1}{4RD}\sum_{n=0}^{N}\bar{X}_{2n+1} \pm \left(1 - \sum_{n=0}^{N}M_{2n+1}\right)\right] \quad \begin{matrix}(+)\text{ if } R > D \\ (-)\text{ if } R < D,\end{matrix} \tag{36}$$

$$\frac{\partial P}{\partial R} \approx R\left[\frac{1}{4RD}\sum_{n=0}^{N}\bar{X}_{2n+1}\right]. \tag{37}$$

The determination of the constant M is discussed in Appendix A of [4]. If the constants M and ϵ were chosen such that

$$M = 30, \qquad \epsilon = 10^{-8}, \tag{38}$$

then sufficient tests were made on the results to assure seven-decimal digit accuracy in the values of P and $\partial P/\partial R$ for all values of R and D. The tests are described in [4]. The maximum number of terms, N', required for seven-decimal

digit accuracy in either series that occurs in equation (27) was twenty for $0 < R \leqq 126$, $0 \leqq D \leqq 120$.

4. Table Computation—Discussion of Results. The extensive inverse table, mentioned in the introduction, has R tabulated as a function of P and D for the

TABLE 1
Inverse $P(R, D)$ Table, $R = R(P, D)$

D / P	0.1	0.5	1.0	1.5	2.0	3.0	4.0	5.0
.01	0.142132	0.150917	0.181965	0.247976	0.377894	0.973968	1.857355	2.807007
.05	0.321093	0.340911	0.410355	0.552995	0.803492	1.589932	2.514287	3.475659
.10	0.460192	0.488541	0.586808	0.780875	1.090931	1.931431	2.867729	3.833372
.15	0.571548	0.606683	0.727145	0.956651	1.299471	2.164629	3.107065	4.075094
.20	0.669719	0.710800	0.850071	1.106744	1.470965	2.351156	3.297689	4.267393
.25	0.760426	0.806964	0.962923	1.241576	1.621141	2.511865	3.461479	4.432486
.30	0.846714	0.898407	1.069594	1.366651	1.757905	2.656649	3.608743	4.580828
.35	0.930528	0.987190	1.172547	1.485396	1.885955	2.791156	3.745340	4.718356
.40	1.013296	1.074827	1.273564	1.600226	2.008448	2.919061	3.875068	4.848912
.45	1.096204	1.162568	1.374100	1.713036	2.127745	3.043037	4.000676	4.975274
.50	1.180355	1.251580	1.475479	1.825472	2.245802	3.165246	4.124378	5.099676
.55	1.266891	1.343064	1.579042	1.939121	2.364426	3.287634	4.248157	5.224119
.60	1.357113	1.438388	1.686286	2.055680	2.485472	3.412162	4.374006	5.350606
.65	1.452637	1.539246	1.799042	2.177146	2.611062	3.541034	4.504154	5.481380
.70	1.555634	1.647914	1.919739	2.306101	2.743883	3.677012	4.641388	5.619238
.75	1.669270	1.767705	2.051892	2.446209	2.887695	3.823927	4.789566	5.768053
.80	1.798604	1.903913	2.201075	2.603222	3.048351	3.987718	4.954663	5.933817
.85	1.952745	2.066052	2.377281	2.787369	3.236215	4.178871	5.147218	6.127099
.90	2.151322	2.274618	2.601947	3.020515	3.473382	4.419704	5.389656	6.370384
.95	2.453851	2.591661	2.939763	3.368463	3.826253	4.777225	5.749279	6.731139
.97	2.654829	2.801806	3.161592	3.595668	4.056141	5.009727	5.982997	6.965523
.99	3.042407	3.205999	3.584494	4.026818	4.491533	5.449368	6.424667	7.408327
.995	3.263342	3.435790	3.823110	4.269216	4.735933	5.695826	6.672133	7.656366
.999	3.726147	3.915765	4.318250	4.770776	5.240984	6.204548	7.182694	8.167991
.9999	4.302554	4.511127	4.927840	5.386401	5.860000	6.827233	7.807274	8.793692
.99999	4.810368	5.033640	5.459903	5.922582	6.398559	7.368429	8.349868	9.337129
.999999	5.269458	5.504595	5.937784	6.403513	6.881283	7.853179	8.835714	9.823646

D / P	6.0	8.0	10.0	20.0	30.0	50.0	80.0	120.0
.01	3.778556	5.747335	7.730490	17.70022	27.69100	47.68389	77.67999	117.6779
.05	4.452164	6.424982	8.409712	18.38123	28.37229	48.36531	78.36146	118.3593
.10	4.811875	6.786445	8.771899	18.74428	28.73548	48.72858	78.72475	118.7226
.15	5.054765	7.030393	9.016299	18.98923	28.98053	48.97367	78.96986	118.9678
.20	5.247904	7.224314	9.210559	19.18391	29.17528	49.16846	79.16466	119.1626
.25	5.413665	7.390705	9.377228	19.35094	29.34237	49.33558	79.33179	119.3297
.30	5.562570	7.540148	9.526912	19.50093	29.49241	49.48565	79.48187	119.4798
.35	5.700590	7.678645	9.665623	19.63992	29.63145	49.62472	79.62094	119.6189
.40	5.831589	7.810077	9.797253	19.77181	29.76339	49.75668	79.75291	119.7508
.45	5.958359	7.937251	9.924613	19.89941	29.89104	49.88435	79.88059	119.8785
.50	6.083144	8.062420	10.04996	20.02499	30.01667	50.01000	80.00625	120.0042
.55	6.207953	8.187598	10.17531	20.15058	30.14229	50.13565	80.13191	120.1298
.60	6.334797	8.314803	10.30269	20.27819	30.26994	50.26332	80.25959	120.2575
.65	6.465923	8.446290	10.43434	20.41008	30.40188	50.39528	80.39156	120.3895
.70	6.604135	8.584868	10.57310	20.54907	30.54092	50.53435	80.53063	120.5286
.75	6.753314	8.734427	10.72284	20.69907	30.69097	50.68442	80.68071	120.6786
.80	6.919464	8.900983	10.88959	20.86610	30.85806	50.85154	80.84784	120.8458
.85	7.113172	9.095143	11.08398	21.06080	31.05282	51.04633	81.04264	121.0406
.90	7.356958	9.339466	11.32857	21.30578	31.29787	51.29143	81.28775	121.2857
.95	7.718391	9.701640	11.69111	21.66887	31.66108	51.65469	81.65104	121.6490
.97	7.953181	9.936878	11.92658	21.90468	31.89696	51.89061	81.88697	121.8849
.99	8.396685	10.38117	12.37128	22.34999	32.34240	52.33612	82.33251	122.3305
.995	8.645082	10.62997	12.62029	22.59934	32.59182	52.58558	82.58198	122.5800
.999	9.157380	11.14304	13.13378	23.11348	33.10609	53.09994	83.09636	123.0943
.9999	9.783802	11.77031	13.76151	23.74194	33.73473	53.72866	83.72513	123.7231
.99999	10.32779	12.31496	14.30652	24.28755	34.28047	54.27449	84.27098	124.2690
.999999	10.81475	12.80245	14.79431	24.77585	34.76889	54.76298	84.75950	124.7575

following ranges:

$$P = 0.01(.01)0.99,$$
$$D = 0(.1)5(.2)10(2)20(5)120,$$

and

$$P = .99(.0005).9990(.0001).9999(.00001).99999(.000001).999999,$$
$$D = 0, .05, .10, .25, .75, 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 20, 30, 50, 80, 120.$$

This table required the calculation of over 45,000 $P(R, D)$ functions to an accuracy of seven or more decimal digits. The tabulated values of R, determined by a Newton-Raphson process, are correct to within one unit in the last digit position given. The method by which this conclusion was verified is given in Appendix C of [4]. A condensed version of the complete table is given below. The complete table as well as a similar one for K as a function of V and c are available by direct request to the authors.

It can be proved that $R(P, D)$ as a function of P approximates a univariate normal distribution to any desired accuracy for sufficiently large fixed values of D and $|R - D|/(R + D) \ll 1$. The relation between R and P in this case is given by

$$P(R, D) \approx \frac{1}{2}\left[1 + \operatorname{Erf}\left(\frac{R - \mu_R}{\sqrt{2}}\right)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{R-\mu_R} e^{-t^2/2}\, dt, \tag{39}$$

where $\mu_R \equiv R(0.50, D) \approx D + 1/(2D)$. (A slightly different formulation of the asymptotic behavior was given by Germond in [7]). This shows that the functional relationship is symmetric with respect to the point $R = \mu_R$, $P = 0.50$. This is evident from a study of Table 1. Also, if $20 \leqq D \leqq 25$, and if μ_R is computed from the approximation $D + 1/(2D)$ (which for these values of D is accurate to 10^{-5} or better), and if values of R as a function of P are then computed from equation (39) by inverse interpolation in an error function or univariate probability integral table, the results are, in general, correct within 10^{-3}, or one unit in the fifth significant figure of R. Further, the accuracy improves rapidly as D increases. This means that an efficient inverse table such as Table 1 need extend only from $P = 0$ to $P = 0.50$ if D is large. Each value of R for $P > 0.50$ is then found with only one subtraction and one addition by using the symmetry property stated above.

5. Acknowledgment. The authors wish to thank Mr. David Eliezer and Mr. Robert Belsky, who programmed and coded the editing procedure for setting up the complete tables, and Mr. Robert Gramp, who programmed and coded the method of computing $V(K, c)$ and $P(R, D)$ for the IBM 7090. The authors are indebted to the referee for suggestions which materially improved the introductory portion of this paper, for correcting a false impression the authors had concerning the origin of equation (9), and for calling the attention of the authors to the univariate normal character of the circular coverage function for large D, as commented on at the end of Section 4.

U. S. Naval Weapons Laboratory
Dahlgren, Virginia

1. H. E. DANIELS, "The covering circle of a sample from a circular normal distribution," *Biometrika*, v. 39, 1952, p. 137–143.

2. N. G. DE BRUIJN, *Asymptotic Methods in Analysis*, North-Holland Publishing Co., Amsterdam, and Interscience Publishers, Inc., New York, 1958.

3. A. R. DIDONATO & M. P. JARNAGIN, "Integration of the general bivariate Gaussian distribution over an offset circle," *Math. Comp.*, v. 15, 1961, p. 375–382.

4. A. R. DIDONATO & M. P. JARNAGIN, *A Method for Computing the Generalized Circular Error Function and the Circular Coverage Function*, NWL Report 1768, U. S. Naval Weapons Laboratory, Dahlgren, Virginia, 23 January 1962.

5. R. V. ESPERTI, *Tables of the Elliptical Normal Probability Function*, Defense Systems Division, General Motors Corporation, Warren, Michigan, 6 April 1960.

6. H. E. FETTIS, *Some Mathematical Identities and Numerical Methods Relating to the Bivariate Normal Probability for Circular Regions*, WADC Technical Note 57-383, ASTIA Document No. AD142135, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio, December, 1957.

7. H. H. GERMOND, *The Circular Coverage Function*, RAND *Corporation Research Memorandum* RM-330, 26 January 1950.

8. A. GRAY, G. B. MATHEWS & T. M. MACROBERT, *A Treatise on Bessel Functions and Their Applications to Physics*, Second Edition, The Macmillan Co., New York and London, 1922.

9. W. C. GUENTHER, "Circular probability problems," *Amer. Math. Monthly*, v. 68, n. 6, 1961, p. 541–544.

10. H. L. HARTER, "Circular error probabilities," *J. Amer. Statist. Assoc.*, v. 55, n. 292, 1960, p. 723–731.

11. J. R. LOWE, "A table of the bivariate normal distribution over an offset circle," *J. Roy. Statist. Soc.*, Ser. B, v. 22, 1960, p. 177–186.

12. *Offset Circle Probabilities*, RAND *Corporation Report R-234*, 14 March 1952.

13. P. B. PATNAIK, "The non-central χ^2- and F-distributions and their applications," *Biometrika*, v. 36, 1949, p. 202–232.

14. *Probability-of-Damage Problems of Frequent Occurrence*, OEG *Study* 626, Operations Evaluation Group, Office of the Chief of Naval Operations, 11 December 1959.

15. H. RUBEN, "Probability content of regions under spherical normal distributions"; I, *Ann. Math. Statist.*, v. 31, 1960, p. 598–618; II, ibid., v. 31, 1960, p. 1113–1121; III, ibid., v. 32, 1961, p. 171–186.

16. H. SOLOMON, *Distribution of Quadratic Forms-Tables and Applications*, Applied Mathematics and Statistics Laboratories Technical Report No. 45, Stanford University, 5 September 1960.

17. *Table of Circular Normal Probabilities, Bell Aircraft Corporation Report* #02-949-106, June 1956. Reviewed in MTAC, v. 11, 1957, p. 210.

18. H. WEINGARTEN & A. R. DIDONATO, "A table of generalized circular error," *Math. Comp.*, v. 15, 1961, p. 169–173.

Optimum-Point Formulas for Osculatory and Hyperosculatory Interpolation

By Herbert E. Salzer

Abstract. Formulas are given for n-point osculatory and hyperosculatory (as well as ordinary) polynomial interpolation for $f(x)$, over $(-1, 1)$, in terms of $f(x_i)$, $f'(x_i)$ and $f''(x_i)$ at the irregularly-spaced Chebyshev points $x_i = -\cos\{(2i-1)\pi/2n\}$, $i = 1, \cdots, n$. The advantage over corresponding formulas for x_i equally spaced is in the squaring and cubing, in the respective osculatory and hyperosculatory formulas, of the approximate ratio of upper bounds for the remainder in ordinary interpolation using Chebyshev and equal spacing (e.g., for $n = 10$, the 15 per cent ratio for ordinary interpolation becoming 2.4 per cent and 0.37 per cent for osculatory and hyperosculatory interpolation). The upper bounds for the remainders in these optimum n-point r-ply confluent formulas (here $r = 1$ and 2) are around 2^r times those of the optimum $\{(r+1)n\}$-point non-confluent formulas. But these present confluent formulas may require fewer computations for irregular arguments when $f(x)$ satisfies a simple first or second-order differential equation. To facilitate computation, for $n = 2(1)10$, auxiliary quantities a_i, b_i and c_i, $i = 1, \cdots, n$, independent of x, are tabulated exactly or to 15S, not precisely for the optimum points, but for those Chebyshev arguments rounded to 2D ("near-optimum" points). At the very worst ($n = 9$, hyperosculatory) this change about doubles the remainder, which is still less than $(\frac{1}{50})$th of the remainder in the corresponding equally-spaced formula.

1. Advantage Over Equal-Interval Formulas. Formulas are given here for n-point osculatory and hyperosculatory polynomial interpolation for $f(x)$, from prescribed values of $f(x)$ with its first, or first and second derivatives at the irregularly-spaced Chebyshev points $x_{n-i+1} = \cos\{(2i-1)\pi/2n\}$, $i = 1, 2, \cdots, n$, instead of equally-spaced points. In this notation, $x_i = -x_{n-i+1}$ and x_i increases with i. For the sake of completeness, the ordinary Lagrangian interpolation formulas are also given for these Chebyshev points. All n-point ordinary, osculatory and hyperosculatory formulas given here are exact for $f(x)$ a polynomial of degree $n-1$, $2n-1$ and $3n-1$ respectively.

The advantage of Chebyshev-point over equal-interval polynomial interpolation formulas is apparent from the factor $\Pi(x) \equiv \Pi_{i=1}^{n}(x - x_i)$ in the remainder term, which is $\Pi(x)f^{(n)}(\xi)/n!$ for n-point ordinary Lagrangian interpolation, $\{\Pi(x)\}^2 f^{(2n)}(\xi)/(2n)!$ for n-point osculatory interpolation and $\{\Pi(x)\}^3 f^{(3n)}(\xi)/(3n)!$ for n-point hyperosculatory interpolation. At the moment, in order to compare Chebyshev-point with equal-interval formulas, let the range of x be $(-1, 1)$, since the relative improvement of the former over the latter is unchanged under any linear transformation. For x_i at the Chebyshev points, $|\Pi(x)| \leqq (\frac{1}{2})^{n-1}$, which is a fraction of the upper bound of $|\Pi(x)|$ for equally-spaced x_i's. However, that fraction is not impressively small, decreasing rather slowly with increasing n (except

Received August 28, 1961.

SCHEDULE 1: *Upper Bound for Absolute Value of Coefficient of* $f^{(m)}(\xi)$

n	Ordinary: $m = n$		Osculatory: $m = 2n$		Hyperosculatory: $m = 3n$	
	U.B.	*Ratio to U.B. for equal spacing*	*U.B.*	*Ratio to U.B. for equal spacing*	*U.B.*	*Ratio to U.B. for equal spacing*
2	.250	50%	.(1)104	25%	.(3)174	12½%
3	.(1)417	65%	.(4)868	42%	.(7)431	27%
4	.(2)521	63%	.(6)388	40%	.(11)408	25%
5	.(3)521	55%	.(8)108	30%	.(15)187	17%
6	.(4)434	45%	.(11)204	20%	.(20)477	9.2%
7	.(5)310	36%	.(14)280	13%	.(25)747	4.5%
8	.(6)194	27%	.(17)292	7.6%	.(30)769	2.1%
9	.(7)108	21%	.(20)238	4.3%	.(35)547	0.90%
10	.(9)538	15%	.(23)157	2.4%	.(40)281	0.37%

for a slight increase from $n = 2$ to $n = 3$) being somewhat larger than $\frac{1}{9}$ for $n = 11$. Thus ordinary Lagrangian interpolation at Chebyshev points, even for $n = 9$ or 10, gains less than one full decimal place accuracy over interpolation at equally-spaced points. But in the osculatory and hyperosculatory cases, the $\{\Pi(x)\}^2$ and $\{\Pi(x)\}^3$ in the remainder term squares and cubes the relative improvement of the Chebyshev-point formulas. For instance, when $n = 10$ the approximately 15 per cent ratio in the upper bounds of $|\Pi(x)|$ for the Chebyshev and equally-spaced points is now replaced by only around 2 per cent and 0.4 per cent in the ratios of the upper bounds of $\{\Pi(x)\}^2$ and $|\{\Pi(x)\}^3|$ respectively.

In Schedule 1, we give the upper bound for the absolute value of the coefficient of $f^{(m)}(\xi)$, $-1 \leqq \xi \leqq 1$, $m = n$, $2n$ and $3n$, in the remainder term of the n-point ordinary, osculatory and hyperosculatory interpolation formulas, for $n = 2(1)10$ to 38. These bounds are, of course, $1/2^{n-1}n!$, $1/2^{2n-2}(2n)!$ and $1/2^{3n-3}(3n)!$ respectively. Next to each upper bound is the ratio, in per cent, of that quantity to the corresponding upper bound when the n points x_i are equally-spaced over $(-1, 1)$. The quantity in parentheses indicates the number of zeros between the decimal point and the first significant digit.

2. Comparison with Non-Osculatory Chebyshev-Point Formulas. The upper bounds for $\{\Pi(x)\}^2$ and $|\{\Pi(x)\}^3|$ in the n-point Chebyshev osculatory and hyperosculatory formulas are $(\frac{1}{2})^{2n-2}$ and $(\frac{1}{2})^{3n-3}$ respectively, which is only twice and four times the upper bounds of $(\frac{1}{2})^{2n-1}$ and $(\frac{1}{2})^{3n-1}$ for $|\Pi(x)|$ in the $2n$- and $3n$-point optimum-point (non-confluent) formulas of the same degree of accuracy, namely, for x_i at the zeros of the Chebyshev polynomials $T_{2n}(x) = (\frac{1}{2})^{2n-1} \cos(2n \cos^{-1} x)$ and $T_{3n}(x) = (\frac{1}{2})^{3n-1} \cos(3n \cos^{-1} x)$. This two-and four-ratio is unchanged, of course, under a linear transformation to any range (a, b) other than $(-1, 1)$, because the factor of $\{(b - a)/2\}^{2n}$ or $\{(b - a)/2\}^{3n}$ which then enters the remainder term is the same for both confluent and non-confluent forms of the interpolation formulas.

The confluent Chebyshev-point formulas given here, while not quite as ac-

curate as the non-confluent Chebyshev-point formulas of the same degree, have this advantage: For irregularly-spaced values of x_i, it is often less work to compute n values of $y_i \equiv f(x_i)$ together with $y_i' \equiv f'(x_i)$, or with y_i' and $y_i'' \equiv f''(x_i)$, instead of $2n$ or $3n$ values of y_i. For instance, in the osculatory case $y = f(x)$ might satisfy a rather simple first-order differential equation $y' = \phi(x, y)$ where it is easier to obtain n values of $y_i' = \phi(x_i, y_i)$ after y_i has been calculated than to compute n more values of y_i. The most obvious example is when $\phi(x, y) = y$, where $y = e^x$ and obtaining $y_i' = y_i$ involves no extra work at all. In the hyperosculatory case y might satisfy a simple second-order differential equation from which y_i'' is readily obtained from y_i and y_i'.

3. Interpolation Formulas. We shall not repeat here the derivations of the interpolation formulas, since they have been given a number of times, as well as a full discussion of their advantages, efficient arrangement, remainder terms, extension to inverse and complex interpolation, etc., in previous articles [1]-[3]. In (1)–(14) below, n is understood, i ranges from 1 to n, $f \equiv f(x)$, $f_i \equiv f(x_i)$, $f_i' \equiv f'(x_i)$, $f_i'' \equiv f''(x_i)$ and $\sum$ denotes $\sum_{i=1}^{n}$. We employ quantities p_{ij}, q_i, r_i and s_i given by

$$(1)\qquad \begin{cases} p_{ij} = 1/(x_i - x_j), j \neq i; & q_i = \displaystyle\sum_{j=1, j\neq i}^{n} p_{ij}; \\ r_i = q_i^2; & s_i = \displaystyle\sum_{j=1, j\neq i}^{n} p_{ij}^2. \end{cases}$$

For each n we define first

$$(2)\qquad A_i = \prod_{j=1, j\neq i}^{n} p_{ij}.$$

For ordinary interpolation we define

$$(3)\qquad a_i = k_1(n) A_i.$$

For osculatory interpolation we define

$$(4)\qquad \begin{cases} a_i = k_2(n) A_i^2, \\ b_i = -2q_i a_i = -2k_2(n) q_i A_i^2. \end{cases}$$

For hyperosculatory interpolation we define

$$(5)\qquad \begin{cases} a_i = k_3(n) A_i^3, \\ b_i = -3q_i a_i = -3k_3(n) q_i A_i^3, \\ c_i = a_i[\tfrac{9}{2} r_i + \tfrac{3}{2} s_i] = k_3(n)[\tfrac{9}{2} r_i + \tfrac{3}{2} s_i] A_i^3. \end{cases}$$

In (3)–(5), the $k_m(n)$, $m = 1, 2, 3$, denote suitably chosen constants that do not affect the results of the interpolation in formulas (7), (10) and (14), but which might (and this depends upon the values and functional nature of the arguments x_i) facilitate appreciably the calculation and use of the auxiliary quantities a_i, a_i and b_i, or a_i, b_i and c_i in (6)–(14).

For ordinary n-point interpolation, of $(n-1)$th degree accuracy, we obtain

$$(6)\qquad \alpha_i = a_i/(x - x_i), \text{ from which}$$

$$(7)\qquad f \sim \Sigma \alpha_i f_i / \Sigma \alpha_i.$$

For n-point polynomial osculatory interpolation of $(2n-1)$th degree accuracy, we obtain

$$\beta_i = a_i/(x - x_i), \tag{8}$$

$$\alpha_i = (\beta_i + b_i)/(x - x_i), \quad \text{from which} \tag{9}$$

$$f \sim \Sigma(\alpha_i f_i + \beta_i f_i')/\Sigma\alpha_i . \tag{10}$$

For n-point polynomial hyperosculatory interpolation of $(3n-1)$th degree accuracy, we obtain

$$\gamma_i = a_i/2(x - x_i), \tag{11}$$

$$\beta_i = (2\gamma_i + b_i)/(x - x_i), \tag{12}$$

$$\alpha_i = (\beta_i + c_i)/(x - x_i), \quad \text{from which} \tag{13}$$

$$f \sim \Sigma(\alpha_i f_i + \beta_i f_i' + \gamma_i f_i'')/\Sigma\alpha_i . \tag{14}$$

4. Use of "Near-Optimum" Points. Instead of taking the x_i precisely equal to the zeros of $T_n(x)$, we now round them off to two decimal places. This makes the osculatory and hyperosculatory formulas "near-optimum" rather than "optimum" point formulas. Three reasons for such a choice are: 1) easier calculation and checking of the table of the auxiliary quantities a_i, b_i and c_i occurring in the interpolation formulas (7), (10) and (14); 2) some of the a_i, for the lower values of n, can be given exactly with much fewer than 15 significant figures; 3) for many functions $f(x)$, it is less work to calculate $f(x_i)$ when x_i is an exact two-decimal argument.

The employment of rounded-off zeros of $T_n(x)$ as the arguments x_i was suggested by Lanczos's use of rounded zeros of Legendre polynomials for a modification of Gaussian quadrature. [4] In this present case, the slight shift in the x_i from exact to rounded Chebyshev points does not produce too great a change in the upper bound for the remainder, (the changes for $n = 7$ and $n = 9$ being appreciably greater than the rest, as seen in Schedule 2). This justifies the terminology "near-optimum", which contrasts sharply with the experience of Lanczos with rounded Gaussian points for quadrature formulas. Thus, quoting his comment on an example [4, p. 410]: "Compared with the Gaussian error, the error has increased by the factor 71, which shows the great sensitivity of the Gaussian method to even small shifts of the zeros." Here, at the worst, for 9-point hyperosculatory interpolation, the choice of the near-optimum instead of optimum points causes the maximum error to be slightly more than doubled. But even then it is less than $(\frac{1}{50})$th of the maximum error in the corresponding equally-spaced formula.

In attempting to estimate the sensitivity in the upper bound of the absolute value of $\Pi(x) = T_n(x)$ for a slight change of Δx_i in every x_i, we differentiate $T_n(x) = \Pi_{i=1}^{n}(x - x_i)$ partially with respect to each x_i, obtaining for $D_n(x)$, the dominant part of the deviation in $\Pi(x)$, the expression

$$D_n(x) = -\sum_{i=1}^{n} \frac{\Pi_{i=1}^{n}(x - x_i)}{x - x_i}\,\Delta x_i , \tag{15}$$

TABLES of a_i, b_i and c_i

			Ordinary Interpolation	Osculatory Interpolation	
n	i	x_i	a_i	a_i	b_i
2	1, 2	∓0.71	∓1	1	±1.40845 07042 2535
3	1, 3	∓0.87	1	1	±3.44827 58620 6897
3	2	0	−2	4	0
4	1, 4	∓0.92	∓1.9	0.361	±2.28481 29567 6948
4	2, 3	∓0.38	±4.6	2.116	±0.98676 86309 79157
5	1, 5	∓0.95	0.3481	0.12117 361	±1.21320 85521 6070
5	2, 4	∓0.59	−0.9025	0.81450 625	±0.67432 28058 42933
5	3	0	1.1088	1.22943 744	0
6	1, 6	∓0.97	∓1.5	0.225	±3.23024 16119 5097
6	2, 5	∓0.71	±4.1	1.681	±2.37511 73717 7847
6	3, 4	∓0.26	∓5.6	3.136	±0.85512 42401 72493
7	1, 7	∓0.97	0.37810 201	0.14296 11299 66040	±2.84410 18250 1904
7	2, 6	∓0.78	−1.04383 446	1.08959 03798 8349	±1.99381 25356 4272
7	3, 5	∓0.43	1.51061 495	2.28195 75271 6350	±1.46095 20152 4591
7	4	0	−1.68976 500	2.85530 57552 2500	0
8	1, 8	∓0.98	∓0.97536 56688	0.09513 38187 87367 1	±2.45239 17122 6555
8	2, 7	∓0.83	±2.81517 71648	0.79252 22469 21137	±2.32927 47108 1650
8	3, 6	∓0.56	∓4.15391 20805	1.72549 85572 5238	±0.93364 45598 54158
8	4, 5	∓0.20	±4.72726 03686	2.23469 90592 5362	∓0.05814 57965 08793 1
9	1, 9	∓0.98	0.46316 76707 68	0.21452 42912 44654	±7.31124 44403 2311
9	2, 8	∓0.87	−1.22781 71566 08	1.50753 49700 6095	±2.70477 66013 7086
9	3, 7	∓0.64	1.82850 70803 23	3.34343 81427 9134	±4.60073 50158 6742
9	4, 6	∓0.34	−2.28761 18102 88	5.23316 77945 6914	±2.44461 28094 1959
9	5	0	2.44750 84316 10	5.99029 75228 0204	0
10	1, 10	∓0.99	∓0.41223 53180 2154	0.16993 79574 24320	±6.73542 63847 8561
10	2, 9	∓0.89	±1.24470 77696 4339	1.54929 74318 1062	±8.10439 99566 8827
10	3, 8	∓0.71	∓1.92977 33728 6557	3.72402 52706 2096	±3.47307 39064 2641
10	4, 7	∓0.45	±2.46258 91833 0950	6.06434 54857 5295	±6.43767 06135 5020
10	5, 6	∓0.16	∓2.73564 36743 5008	7.48374 63130 1161	∓1.57793 67089 3871

Hyperosculatory Interpolation

n	i	x_i	a_i	b_i	c_i
2	1, 2	∓0.71	∓1	−2.11267 60563 3803	∓2.97560 00793 4934
3	1, 3	∓0.87	1	±5.17241 37931 0345	15.85414 18945 700
3	2	0	−8	0	−31.70828 37891 399
4	1, 4	∓0.92	∓0.6859	−6.51171 69267 9301	∓35.35105 61593 570
4	2, 3	∓0.38	±9.7336	6.80870 35537 5619	±86.36831 12988 727
5	1, 5	∓0.95	0.04218 05336 41	±0.63347 68455 10712	5.35936 14929 5057
5	2, 4	∓0.59	−0.73509 18906 25	∓0.91286 44984 09870	−13.49924 33461 453
5	3	0	1.36320 02334 72	0	16.27976 37063 895
6	1, 6	∓0.97	∓0.03375	−0.72680 43626 88967	∓8.74001 29929 3812
6	2, 5	∓0.71	±0.68921	1.46069 71836 4376	±23.92400 98678 306
6	3, 4	∓0.26	∓1.75616	−0.71830 43617 44894	∓32.66402 28607 687
7	1, 7	∓0.97	0.00540 53890 59203 10	±0.16130 40925 02655	2.67672 79073 0026
7	2, 6	∓0.78	−0.11373 51985 80688	∓0.31218 15347 22577	−7.06970 10338 4325
7	3, 5	∓0.43	0.34471 59155 79822	±0.33104 03933 19465	10.26580 84638 704
7	4	0	−0.48247 95729 47777	0	−11.74567 06746 548

TABLES of a_i, b_i and c_i—(*Continued*)

Hyperosculatory Interpolation

n	i	x_i	a_i	b_i	c_i
8	1, 8	∓0.98	∓0.00927 90260 78703 83	−0.35879 68023 89020	∓7.68100 77077 0057
8	2, 7	∓0.83	±0.22310 90532 12837	0.98359 81464 65512	±23.18826 27433 073
8	3, 6	∓0.56	∓0.71675 69301 85600	−0.58174 16124 10694	∓33.10310 66120 670
8	4, 5	∓0.20	±1.05640 04298 5573	−0.04123 05479 15504 7	±34.10349 18155 463
9	1, 9	∓0.98	0.00993 60716 29894 27	±0.50794 98086 75992	14.41869 83299 256
9	2, 8	∓0.87	−0.18509 77300 42737	∓0.49814 56673 93253	−30.70431 01740 749
9	3, 7	∓0.64	0.61135 00316 71595	±1.26187 14826 8053	41.26425 87109 550
9	4, 6	∓0.34	−1.19714 56452 0752	∓0.83884 87701 61438	−54.56736 40851 665
9	5	0	1.46613 03694 9105	0	59.17743 44367 215
10	1, 10	∓0.99	∓0.00700 54427 92274 56	−0.41648 70956 61415	∓13.63892 25685 320
10	2, 9	∓0.89	±0.19284 22550 86323	1.51314 14391 5812	±46.53720 81301 798
10	3, 8	∓0.71	∓0.71865 24807 12282	−1.00533 68319 9238	∓70.78787 62690 287
10	4, 7	∓0.45	±1.49339 91597 0670	2.37800 07027 9572	±93.69814 23293 455
10	5, 6	∓0.16	∓2.04728 63261 6309	0.64750 08864 49945	∓104.39709 18091 15

so that

$$|D_n(x)| \leqq 2^{-n+1} \sum_{i=1}^{n} \frac{|\Delta x_i|}{|x - x_i|}. \tag{16}$$

Now for x in the neighborhood of the extrema of $T_n(x)$ not close to the ends ± 1, the $|x - x_i|$ stays large enough for (16) to furnish upper bounds for $|D_n(x)|/2^{-n+1}$ of the order of just several per cent when Δx_i is the roundoff error in employing x_i to 2D. However (16) breaks down as a practical formula, for larger n and x either at ± 1 or at an extremum close to ± 1 since there $|x - x_i|$ is quite small. This might also be expected from the very large derivative of $2^{n-1}T_n(x)$ at $x = \pm 1$, its magnitude being n^2. Thus, to be on the safe side, to provide for every x in the range $(-1, 1)$, instead of using (15) or (16), the factor $\Pi(x) = \Pi_{i=1}^{n}(x - x_i)$ for the chosen near-optimum x_i's was calculated for every n from 2 to 10, for $x = -1(.001)1$, and its greatest deviation from zero was found. The percentage increase in the upper bound for the absolute value of the coefficient of $f^{(m)}(\xi)$ (see Schedule 1), due to the use of these near-optimum points x_i instead of optimum points, is given in Schedule 2.

SCHEDULE 2: *Increase in Schedule 1 When Using Near-Optimum Points*

n	Ordinary	Osculatory	Hyperosculatory
2	0.82%	1.65%	2.5%
3	1.4%	2.8%	4.2%
4	5.1%	10.5%	16%
5	1.7%	3.4%	5.2%
6	2.6%	5.3%	8.0%
7	21%	46%	76%
8	6.2%	13%	20%
9	29%	66%	113%
10	7.6%	16%	25%

5. Tables of Auxiliary Coefficients a_i, b_i and c_i. To facilitate the use of (6)–(14) for these near-optimum points x_i, the auxiliary quantities a_i, b_i and c_i are tabulated here for $n = 2(1)10$, $i = 1, \cdots, n$. It reduced the work considerably to choose the constants $k_m(n)$, $m = 1, 2, 3$, in (3)–(5), as products of powers of selected prime numbers < 200. As a result of this choice, it was easy to give exact values of all the quantities a_i for ordinary interpolation, and of a_i for $n = 2(1)6$ for osculatory and hyperosculatory interpolation. The remaining quantities a_i and all quantities b_i and c_i are given to 15S, believed to be correct to within a unit in the last place. In reading entries prefixed by $\pm$ or $\mp$ signs, the upper sign corresponds to the negative x_i.

General Dynamics/Astronautics
San Diego, California

1. H. E. Salzer, "New formulas for facilitating osculatory interpolation," *J. Res. Nat. Bur. Standards*, v. 52, 1954, p. 211–216.
2. H. E. Salzer, "Formulae for hyperosculatory interpolation, direct and inverse," *Quart. J. Mech. Appl. Math.*, v. 12, 1959, p. 100–110.
3. H. E. Salzer, "Alternative formulas for osculatory and hyperosculatory inverse interpolation," *Math. Comp.*, v. 14, 1960, p. 257–261.
4. C. Lanczos, *Applied Analysis*, Prentice Hall, Englewood Cliffs, New Jersey, 1956, p. 408–410, table on p. 529.

A Heuristic Asymptotic Formula Concerning the Distribution of Prime Numbers

By Paul T. Bateman and Roger A. Horn

Suppose $f_1, f_2, \cdots, f_k$ are polynomials in one variable with all coefficients integral and leading coefficients positive, their degrees being $h_1, h_2, \cdots, h_k$ respectively. Suppose each of these polynomials is irreducible over the field of rational numbers and no two of them differ by a constant factor. Let $Q(f_1, f_2, \cdots, f_k; N)$ denote the number of positive integers n between 1 and N inclusive such that $f_1(n), f_2(n), \cdots, f_k(n)$ are all primes. (We ignore the finitely many values of n for which some $f_i(n)$ is negative.) Then heuristically we would expect to have for N large

$$(1) \qquad Q(f_1, \cdots, f_k; N) \sim h_1^{-1}h_2^{-1} \cdots h_k^{-1} C(f_1, \cdots, f_k) \int_2^N (\log u)^{-k}\, du,$$

where

$$C(f_1, f_2, \cdots, f_k) = \prod_p \left\{\left(1 - \frac{1}{p}\right)^{-k} \left(1 - \frac{\omega(p)}{p}\right)\right\},$$

the product being taken over all primes and $\omega(p)$ being the number of solutions of the congruence

$$f_1(x)\, f_2(x) \cdots f_k(x) \equiv 0 \pmod{p}.$$

(If $\omega(p) = p$ for some prime p, then $C(f_1, f_2, \cdots, f_k) = 0$ and $Q(f_1, f_2, \cdots, f_k) \leqq h_1 + h_2 + \cdots + h_k$ for all N; we agree to exclude this trivial case.) Although it does not seem likely that (1) will be proved in the foreseeable future (aside from the known case of a single linear polynomial), a simple application of Atle Selberg's sieve method [1] does show that

$$(2) \quad Q(f_1, \cdots, f_k; N) \leqq 2^k k!\, C(f_1, \cdots, f_k) \int_2^N (\log u)^{-k}\, du + o(N(\log N)^{-k}).$$

The details of the argument were given by Bateman and Stemmler [2]. Thus we at least know an upper bound for $Q(f_1, f_2, \cdots, f_k; N)$ which is $2^k\, k!\, h_1h_2 \cdots h_k$ times the conjectured asymptotic value.

The heuristic argument in support of (1) may be put into various forms, but essentially amounts to the following. In some sense the chance that a large positive integer m is prime is around $1/\log m$. Since $\log f_i(n)$ is around $h_i \log n$, the chance that $f_1(n), f_2(n), \cdots, f_k(n)$ are all primes would seem to be about

$$h_1^{-1}h_2^{-1} \cdots h_k^{-1} (\log n)^{-k}.$$

Received November 20, 1961. This work was supported by the National Science Foundation and the Office of Naval Research. The authors would like to thank Professor Atle Selberg for some helpful remarks.

But this ignores the fact that $f_1(n), f_2(n), \cdots, f_k(n)$ are not quite random integers. Thus for each prime p we must apply a correction factor r_p/s_p, where r_p is the chance that for random n none of the integers $f_1(n), f_2(n), \cdots, f_k(n)$ is divisible by p and s_p is the chance that none of the integers in a random k-tuple is divisible by p. But clearly $r_p = 1 - \omega(p)/p$ and $s_p = (1 - 1/p)^k$. Thus the chance that $f_1(n), f_2(n), \cdots, f_k(n)$ are all primes for a large positive integer n taken at random is about

$$h_1^{-1} h_2^{-1} \cdots h_k^{-1} C(f_1, \cdots, f_k) (\log n)^{-k}.$$

Hence we would expect $Q(f_1, f_2, \cdots, f_k; N)$ to be about

$$h_1^{-1} h_2^{-1} \cdots h_k^{-1} C(f_1, \cdots, f_k) \sum_{n=2}^{N} (\log n)^{-k},$$

which is essentially the same as the approximation given in (1). This "derivation" of the heuristic formula (1) by the simple expedient of multiplying p-adic densities has the virtue that an analogous method gives results known to be correct in Waring's problem and in the theory of quadratic forms.

The convergence of the product defining $C(f_1, \cdots, f_k)$ may be proved as follows. Let $\omega_i(p)$ be the number of solutions of the congruence

$$f_i(x) \equiv 0 \pmod{p}.$$

If $i \neq j$, there exist polynomials g_{ij} and h_{ij} with integral coefficients and a positive integer c_{ij} such that

$$g_{ij}(x) f_i(x) + h_{ij}(x) f_j(x) = c_{ij}$$

identically in x. Thus

$$\omega(p) = \omega_1(p) + \cdots + \omega_k(p)$$

for all sufficiently large primes p, namely for all primes p such that

$$p \nmid \prod_{1 \leq i < j \leq k} c_{ij}.$$

For all but a finite number of primes p we also know that $\omega_i(p)$ is the number of distinct prime ideals of norm p in the algebraic number field generated by a zero of f_i. Hence by an elementary result on the distribution of prime ideals we have for x large

$$\sum_{p \leq x} \frac{\omega_i(p)}{p} = \log \log x + A_i + o(1)$$

$$= \sum_{p \leq x} \frac{1}{p} + B_i + o(1),$$

where A_i and B_i are certain constants. Thus the series

$$\sum_p \frac{\omega_i(p) - 1}{p}$$

converges for $i = 1, \cdots, k$ and so

$$\sum_p \frac{\omega(p) - k}{p}$$

converges. Since

$$\left(1-\frac{1}{p}\right)^{-k}\left(1-\frac{\omega(p)}{p}\right)=1+\frac{k-\omega(p)}{p}+\frac{B(p)}{p^2},$$

where $B(p)$ is a bounded function of p, it follows that the product for $C(f_1, \cdots, f_k)$ converges.

Numerical evidence has been given for many special cases of (1), particularly when the f_i are all linear polynomials [3], [4]. The case of the single polynomial n^2+1 was discussed by Western [5] and other polynomials of the form n^2+a were considered by Shanks [6]. The case of the pair of polynomials n^2-2n+2 and n^2+2n+2 was discussed by Shanks [7]. The case of the single polynomial n^4+1 was treated by Shanks [8]. In all cases the heuristic formula (1) fits the numerical data remarkably well.

In this note we give some numerical data for the case of the pair of polynomials n and n^2+n+1. That is, we consider the number $P(N)$ of primes p not exceeding N such that p^2+p+1 is also a prime. The heuristic formula here is

$$P(N)\sim\frac{1}{2}C\int_2^N(\log u)^{-2}\,du, \tag{3}$$

where

$$C=\prod_p\left\{\left(1-\frac{1}{p}\right)^{-2}\left(1-\frac{2+\chi(p)}{p}\right)\right\},$$

where $\chi(p)$ is -1, 0, or $+1$ according as p is congruent to -1, 0, or $+1$ modulo 3. As in every case where each of the polynomials f_i has the property that a zero thereof generates a normal extension of the field of rational numbers with an abelian galois group, we can express C in terms of an absolutely convergent infinite product. Specifically

$$C=L(1,\chi)^{-1}\prod_p\left\{1-\frac{p+2p\chi(p)-\chi(p)}{(p-1)^2(p-\chi(p))}\right\},$$

where

$$L(1,\chi)=\prod_p\left(1-\frac{\chi(p)}{p}\right)^{-1}=\sum_{n=1}^{\infty}\frac{\chi(n)}{n}=\frac{\pi}{3\sqrt{3}}.$$

The numerical value of C is about 1.522.

The second-named author used the ILLIAC to prepare a list of the 776 primes of the form p^2+p+1 with p a prime less than 113,000. (The program used was a straightforward one, and the running time was about 400 minutes.) The first 209 of these primes are listed by Bateman and Stemmler [2] who considered primes of the form p^2+p+1 in connection with a problem in algebraic number theory. The table in the present paper gives a comparison of the actual count of $P(N)$ with the value obtained from the right-hand side of (3), rounded off to the nearest integer.

TABLE

N	$P(N)$	$0.761 \int_2^N (\log u)^{-2}\, du$
10	3	3
100	8	8
1000	23	26
5000	71	75
10000	117	123
20000	206	206
30000	275	281
40000	341	350
50000	404	416
60000	469	480
70000	532	542
80000	595	603
90000	649	662
100000	706	720
110000	758	777
113000	776	794

Needless to say, it would probably be worthwhile to attempt some numerical work in cases where at least one of the polynomials $f_1, f_2, \cdots, f_k$ has the property that a zero thereof does *not* generate an abelian extension of the field of rational numbers. An interesting example of this type would be the case of a single polynomial of the form $n^3 + a$, where a is an integer which is not a perfect cube. Of course the determination of the numerical value of $C(f_1, f_2, \cdots, f_k)$ would be much more difficult in such a case.

It should be mentioned that the conjecture (1) and the result (2) may easily be extended to the case where $f_1, f_2, \cdots, f_k$ are merely integral-valued polynomials, that is, polynomials with rational coefficients which take integral values for all integral values of the variable. (For example, $\frac{1}{2}n^2 + \frac{1}{2}n + 1$ is such a polynomial.) The only change that must be made is that $\omega(p)/p$ in the formula for $C(f_1, f_2, \cdots, f_k)$ must be replaced by something a little more complicated, as we now explain. If p is a prime and m is a positive integer, let $\omega(p, m)$ denote the number of integers j between 1 and p^m inclusive such that $f_1(j)\, f_2(j) \cdots f_k(j)$ is a multiple of p. For given p, the ratio $\omega(p, m)/p^m$ has a value $\rho(p)$ independent of m when m is sufficiently large (specifically, when p^m is larger than the highest power of p dividing the denominator of any coefficient in any of the polynomials f_i). We need only replace $\omega(p)/p$ in the formula for $C(f_1, f_2, \cdots, f_k)$ by $\rho(p)$ in order to generalize (1) and (2) to arbitrary integral-valued polynomials. In the case of either (1) or (2), the generalization is an immediate consequence of the special case given originally.

Finally we remark that the conjectural formula (1) may be regarded as a quantitative form of the Hypothesis H of A. Schinzel [9], [10].

University of Illinois
Urbana, Illinois

1. ATLE SELBERG, "On an elementary method in the theory of primes," *Norske Vid. Selsk. Forh. Trondheim*, v. 19, no. 18, 1947, p. 64–67.
2. PAUL T. BATEMAN & ROSEMARIE M. STEMMLER, "Waring's problem in algebraic number fields and primes of the form $(p^r - 1)/(p^d - 1)$," *Illinois J. Math.*, v. 6, 1962, p. 142–156.
3. G. H. HARDY & J. E. LITTLEWOOD, "Some problems of partitio numerorum; III: On the expression of a number as a sum of primes," *Acta Math.*, v. 44, 1923, p. 1–70.
4. D. H. LEHMER, "Tables concerning the distribution of primes up to 37 millions," 1957, deposited in the UMT file and reviewed in *MTAC* v. 13, 1959, p. 56–57.
5. A. E. WESTERN, "Note on the number of primes of the form $n^3 + 1$," *Proc. Cambridge Philos. Soc.*, v. 21, 1922, p. 108–109.
6. DANIEL SHANKS, "On the conjecture of Hardy and Littlewood concerning the number of primes of the form $n^2 + a$," *Math. Comp.*, v. 14, 1960, p. 321–332.
7. DANIEL SHANKS, "A note on Gaussian twin primes," *Math. Comp.*, v. 14, 1960, p. 201–203.
8. DANIEL SHANKS, "On numbers of the form $n^4 + 1$," *Math. Comp.*, v. 15, 1961, p. 186–189.
9. A. SCHINZEL & W. SIERPIŃSKI, "Sur certaines hypothèses concernant les nombres premiers," *Acta Arith.*, v. 4, 1958, p. 185–208.
10. A. SCHINZEL, "Remarks on the paper 'Sur certaines hypothèses concernant les nombres premiers'," *Acta Arith.*, v. 7, 1961, p. 1–8.

Primitive Polynomials (Mod 2)

By E. J. Watson

The following list contains one example of a primitive polynomial (mod 2) for each degree n, $1 \leqq n \leqq 100$. It was compiled with the aid of the Mercury computer at Manchester University by the following method.

The polynomials $P_n(x)$ (mod 2) of degree n were tested in their natural order until a primitive polynomial was found. The test comprised three stages. In the first stage the small primes, of degree up to 9, were tried as possible factors (mod 2) of P_n. If no factor was found P_n went forward to the second stage, which tested whether P_n divides $x^N - 1$, where $N = 2^n - 1$. If it does, and N is prime (a Mersenne prime), this suffices to prove that P_n is primitive. If N is composite, however, P_n might divide $x^M - 1$, where M is a factor of N, and then P_n would not be primitive. The third stage was, therefore, a trial of this possibility, in which M took the values N/p, where p runs through the prime factors of N.

The two latter stages were carried out by a process in which the computer repeated the operations of squaring, possibly multiplying by x (depending on the binary representation of M), then dividing by P_n. The prime factors of N were taken from the tables of Kraïtchik [1], supplemented by Robinson's [2] further decomposition of $2^{95} - 1$. If any more of these 'prime' factors should turn out to be composite, doubt would be cast on the corresponding P_n. Mersenne polynomials for $n = 107$ and 127 are also given. The prime $x^{127} + x + 1$ was found by Zierler [3]. Its nature follows from the general result that if $\Sigma a_n x^n$ divides $\Sigma c_n x^n$ (mod p), then

$$\Sigma a_n x^{p^n} \quad \text{divides} \quad \Sigma c_n x^{p^n} \qquad (\text{mod } p).$$

The primitive character of each polynomial $P_n(x)$ listed has been checked by a repetition of the second and third stages on the conjugate polynomial $x^n P_n(x^{-1})$. In the list only the degrees of the separate terms in P_n are given, thus

$$127 \quad 1 \quad 0 \qquad \text{stands for} \qquad x^{127} + x + 1.$$

Department of Mathematics
University of Manchester

1. M. Kraïtchik, *Introduction à la Théorie des Nombres*, Gauthier-Villars, Paris, 1952.
2. R. M. Robinson, "Some factorizations of numbers of the form $2^n \pm 1$," *MTAC*, v. 11, 1957, p. 265–268.
3. N. Zierler, "Linear recurring sequences," *J. Soc. Indust. Appl. Math.*, v. 7, 1959, p. 31–48.

Received December 18, 1961.

Primitive Polynomials (mod 2)

1	0						51	6	3	1	0			
2	1	0					52	3	0					
3	1	0					53	6	2	1	0			
4	1	0					54	6	5	4	3	2	0	
5	2	0					55	6	2	1	0			
6	1	0					56	7	4	2	0			
7	1	0					57	5	3	2	0			
8	4	3	2	0			58	6	5	1	0			
9	4	0					59	6	5	4	3	1	0	
10	3	0					60	1	0					
11	2	0					61	5	2	1	0			
12	6	4	1	0			62	6	5	3	0			
13	4	3	1	0			63	1	0					
14	5	3	1	0			64	4	3	1	0			
15	1	0					65	4	3	1	0			
16	5	3	2	0			66	8	6	5	3	2	0	
17	3	0					67	5	2	1	0			
18	5	2	1	0			68	7	5	1	0			
19	5	2	1	0			69	6	5	2	0			
20	3	0					70	5	3	1	0			
21	2	0					71	5	3	1	0			
22	1	0					72	6	4	3	2	1	0	
23	5	0					73	4	3	2	0			
24	4	3	1	0			74	7	4	3	0			
25	3	0					75	6	3	1	0			
26	6	2	1	0			76	5	4	2	0			
27	5	2	1	0			77	6	5	2	0			
28	3	0					78	7	2	1	0			
29	2	0					79	4	3	2	0			
30	6	4	1	0			80	7	5	3	2	1	0	
31	3	0					81	4	0					
32	7	5	3	2	1	0	82	8	7	6	4	1	0	
33	6	4	1	0			83	7	4	2	0			
34	7	6	5	2	1	0	84	8	7	5	3	1	0	
35	2	0					85	8	2	1	0			
36	6	5	4	2	1	0	86	6	5	2	0			
37	5	4	3	2	1	0	87	7	5	1	0			
38	6	5	1	0			88	8	5	4	3	1	0	
39	4	0					89	6	5	3	0			
40	5	4	3	0			90	5	3	2	0			
41	3	0					91	7	6	5	3	2	0	
42	5	4	3	2	1	0	92	6	5	2	0			
43	6	4	3	0			93	2	0					
44	6	5	2	0			94	6	5	1	0			
45	4	3	1	0			95	6	5	4	2	1	0	
46	8	5	3	2	1	0	96	7	6	4	3	2	0	
47	5	0					97	6	0					
48	7	5	4	2	1	0	98	7	4	3	2	1	0	
49	6	5	4	0			99	7	5	4	0			
50	4	3	2	0			100	8	7	2	0			
107	7	5	3	2	1	0	127	1	0					

Note on Latent Roots and Vectors of Segments of the Hilbert Matrix

By H. H. Denman and R. C. W. Ettinger

As a by-product of work on condition numbers, maximum and minimum latent roots, and the corresponding vectors, have been calculated for the Hilbert segments of orders 4, 6, 8, and 10, to about 17 significant digits.

This supplements the data of Fairthorne and Miller [1] by including minimum latent roots and corresponding vectors, and increasing the accuracy. It also verifies their data (after rounding) except that one error was found (see next paragraph) in their published results.

A Hilbert segment of order N is the matrix $\| 1/(m+n-1) \|; m, n = 1, \cdots, N$. Let H represent a Hilbert segment, T its inverse, λ a latent root, and V a corresponding vector. Then, e.g., $\lambda_1(H_6)$ is the largest latent root of the segment of order 6; $V_N(T_6)$ is the vector corresponding to the smallest latent root of the inverse of H_6. The error in Fairthorne and Miller's article occurs in the third element of $V_1(H_6)$.

The power method [2] was used, with a double-precision floating-point routine on the IBM 650. Smallest latent roots were obtained as $\lambda_N(H_N) = 1/\lambda_1(T_N)$, and verified (to as many significant figures as the method allows) by direct calculation of $\lambda_N(H_N)$ by the power method, $\lambda_N(H)$ being obtained as $\lambda_1(H - pI)$, where I is the identity matrix and p is slightly greater than $\lambda_1(H)$. The Nth vectors are given as $V_N(H_N) = V_1(T_N)$, because the method gives greater accuracy here for V_1 than for V_N and because T has no input error.

Terminal digits are uncertain by not more than one, as indicated by convergence rates.

$\lambda_1(H_4)$

1. 50021 42800 59242 81

$V_1(H_4)$

1.
0. 57017 20836 63235 83
0. 40677 89880 27529 24
0. 31814 09688 73793 96

$\lambda_1(H_6)$

1. 61889 98589 24339 1

$V_1(H_6)$

1
0. 58862 85434 25543 2
0. 42832 72844 28956 1
0. 33966 18918 38709 5
0. 28252 35879 42149 2
0. 24233 78111 22849 5

$\lambda_1(H_8)$

1. 69593 89969 21949 4

$\lambda_N(H_4)$

$10^{-4} \times$ 0. 96702 30402 25868 861

$V_N(H_4)$

0. 03688 76826 14141 047
−0. 41534 92877 80311 17
1
−0. 65017 12197 33679 82

$\lambda_N(H_6)$

$10^{-7} \times$ 1. 08279 94845 65549 8

$V_N(H_6)$

0. 00180 94825 41440 515
−0. 05161 82535 94248 58
0. 34890 77525 35503 9
−0. 90671 76845 78412 7
1
−0. 39374 11114 93702 0

$\lambda_N(H_8)$

$10^{-10} \times$ 1. 11153 89663 72442 4

Received July 17, 1961.

$V_1(H_8)$ | $V_N(H_8)$

1. | −0. 00006 86103 92145 12812
0. 60050 42457 57953 8 | 0. 00368 78770 51827 661
0. 44267 15540 11918 6 | −0. 04826 72545 24498 43
0. 35437 04469 99697 8 | 0. 26171 33996 76104 1
0. 29691 85784 44507 1 | −0. 70574 73471 79618 8
0. 25618 09294 86980 5 | 1.
0. 22562 93688 08227 6 | −0. 71250 91381 80124 8
0. 20179 01870 37918 3 | 0. 20124 18343 83776 4

$\lambda_1(H_{10})$ | $\lambda_N(H_{10})$

1. 75191 96702 65177 5 | $10^{-13} \times$ 1. 09315 38198 57659 9

$V_1(H_{10})$ | $V_N(H_{10})$

1. | 0. 00000 27147 13133 60409 7
0. 60899 19143 69650 3 | −0. 00023 60612 62959 0383
0. 45313 82989 59421 5 | 0. 00505 28973 86716 890
0. 36528 60134 02151 0 | −0. 04611 60400 49989 25
0. 30775 30474 45501 6 | 0. 22066 15177 28910 4
0. 26672 51842 93050 8 | −0. 60817 67839 54336 8
0. 23580 13079 82484 3 | 1.
0. 21156 39639 51540 1 | −0. 96815 88795 12219 1
0. 19200 51281 86119 1 | 0. 50907 35851 67138 3
0. 17586 00343 93102 9 | −0. 11210 49402 14747 4

Wayne State University
Detroit 2, Michigan

1. R. A. Fairthorne & J. C. P. Miller, "Hilbert's double series theorem and principal latent roots of the resulting matrix," *MTAC*, v. 3, 1949, p. 399.
2. Marvin Marcus, "Basic theorems in matrix theory," *Nat. Bur. Standards, Appl. Math. Ser.* No. 57, U. S. Government Printing Office, Washington, D. C., 1960.

Expansions of Jacobian Elliptic Functions in Powers of the Modulus

By Y. K. Lin* and F. A. Lee

The Maclaurin expansions for Jacobian elliptic functions in the neighborhood of zero modulus (to be called the k-expansions) are given herein. Since the absolute value of modulus k can be taken as smaller than one without loss of generality, the k expansions in most cases converge faster than similar expansions in the neighborhood of the zero argument [1] (to be called the u-expansions). The relative accuracies of the two types of expansions will be illustrated numerically for several pairs of k and u values and by taking equal numbers of terms in the computation. The k expansions also provide for the information, in a fundamental way, as to how the Jacobian functions vary with the change of modulus. The information may be useful in the study of damped non-linear vibrations.

The basic Jacobian function $sn(u, k^2)$ is related to its argument, u, and modulus, k, as the following:

$$(1) \qquad u = \int_0^{sn(u,k)} \frac{dt}{\sqrt{1-t^2}\sqrt{1-k^2t^2}} = \int_0^{sn(u,\phi)} \frac{dt}{\sqrt{1-t^2}\sqrt{1-\phi t^2}}.$$

By successive differentiations one obtains

$$(2) \quad \begin{cases} \dfrac{\partial sn}{\partial \phi} = -cn\cdot dn \displaystyle\int_0^{sn} \frac{(t^2/2)\,dt}{\sqrt{(1-t^2)(1-\phi t^2)^3}} \\ \dfrac{\partial^2 sn}{\partial \phi^2} = -cn\cdot dn \left[\displaystyle\int_0^{sn} \frac{1\cdot 3(t^2/2)^2\,dt}{\sqrt{(1-t^2)(1-\phi t^2)^5}} + \frac{sn^2}{2cn\cdot dn^3}\cdot\frac{\partial sn}{\partial \phi} + \frac{\partial sn}{\partial \phi}\cdot\frac{\partial}{\partial \phi}\left(\frac{1}{cn\cdot dn}\right)\right] \\ \qquad \cdots \\ \dfrac{\partial^m sn}{\partial \phi^m} = -cn\cdot dn \left\{ \displaystyle\int_0^{sn} \frac{1\cdot 3\cdot 5\cdots(2m-1)(t^2/2)^m}{\sqrt{(1-t^2)(1-\phi t^2)^{2m+1}}}\,dt \right. \\ \quad + \dfrac{\partial sn}{\partial \phi}\left[\frac{1\cdot 3\cdot 5\cdots(2m-3)(sn^2/2)^{m-1}}{cn\cdot dn^{2m-1}} \right. \\ \quad + \dfrac{\partial}{\partial \phi}\left(\frac{1\cdot 3\cdot 5\cdots(2m-5)(sn^2/2)^{m-2}}{cn\cdot dn^{2m-3}}\right) + \cdots + \left.\frac{\partial^{m-1}}{\partial \phi^{m-1}}\left(\frac{1}{cn\cdot dn}\right)\right] \\ \quad + \dfrac{\partial^2 sn}{\partial \phi^2}\left[\frac{1\cdot 3\cdot 5\cdots(2m-5)(sn^2/2)^{m-2}}{cn\cdot dn^{2m-3}}\right. \\ \quad + 2\dfrac{\partial}{\partial \phi}\left(\frac{1\cdot 3\cdot 5\cdots(2m-7)(sn^2/2)^{m-3}}{cn\cdot dn^{2m-5}}\right) \\ \quad + \cdots + (m-1)\left.\frac{\partial^{m-2}}{\partial \phi^{m-2}}\left(\frac{1}{cn\cdot dn}\right)\right] + \cdots \\ \quad \left. + \dfrac{\partial^{m-1} sn}{\partial \phi^{m-1}}\left[(m-1)\dfrac{\partial}{\partial \phi}\left(\frac{1}{cn\cdot dn}\right)\right]\right\}. \end{cases}$$

Received July 24, 1961.

* Consultant, and Assistant Professor of Aeronautical Engineering, University of Illinois.

It follows upon evaluating these derivatives at $\phi = 0$ and noting that $\phi = k^2$,

$$sn(u, k^2) = \sin u - \frac{\cos u}{4}\left(u - \frac{\sin 2u}{2}\right)k^2$$

$$(3) \qquad - \frac{\cos u}{4}\left[\frac{3}{4}\left(\frac{3u}{2} - \sin 2u + \frac{\sin 4u}{8}\right) - \sin^2 u\left(u - \frac{\sin 2u}{2}\right)\right.$$

$$\left. + \frac{\tan u}{4}\left(u - \frac{\sin 2u}{2}\right)^2\right]\frac{k^4}{2!} - \cdots + \left\{\frac{\partial^m sn}{\partial \phi^m}\right\}_{\phi=0} \cdot \frac{k^{2m}}{m!} - \cdots.$$

Expansions for other Jacobian functions can be generated from the relations between these functions and sn. The results are given explicitly up to k^4 in the following:

$$cn(u, k^2) = \cos u + \frac{\sin u}{4}\left(u - \frac{\sin 2u}{2}\right)k^2 - \left\{\frac{\cos u}{16}\left(u - \frac{\sin 2u}{2}\right)^2\right.$$

$$(4) \qquad \left. + \frac{\sin u}{4}\left[\sin^2 u\left(u - \frac{\sin 2u}{2}\right) - \frac{3}{4}\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right)\right]\right\}\frac{k^4}{2!} + \cdots.$$

$$(5) \qquad dn(u, k^2) = 1 - \frac{1}{2}(\sin^2 u)k^2 + \left(\frac{u \sin 2u}{4} - \frac{\sin^2 2u}{8} - \frac{1}{4}\sin^4 u\right)\frac{k^4}{2!} + \cdots.$$

$$ns(u, k^2) = \csc u + \csc^2 u\left(\frac{\cos u}{4}\right)\left(u - \frac{\sin 2u}{2}\right)k^2 + \left[-\frac{\cos u}{4}\left(u - \frac{\sin 2u}{2}\right)\right.$$

$$(6) \qquad - \frac{3}{16}\cos u \csc^2 u\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right)$$

$$\left. + \frac{1}{16}\csc u(\csc^2 u + \cot^2 u)\left(u - \frac{\sin 2u}{2}\right)^2\right]\frac{k^4}{2!} + \cdots.$$

$$cs(u, k^2) = \cot u + \frac{\csc^2 u}{4}\left(u - \frac{\sin 2u}{2}\right)k^2 + \left[-\frac{1}{8}\left(u - \frac{\sin 2u}{2}\right)^2 \cot u \csc^2 u\right.$$

$$(7) \qquad \left. + \frac{3}{16}\csc^2 u\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right) - \frac{1}{4}\left(u - \frac{\sin 2u}{2}\right)\right]\frac{k^4}{2!} + \cdots.$$

$$ds(u, k^2) = \csc u + \left[\frac{1}{4}\cot u \csc u\left(u - \frac{\sin 2u}{2}\right) - \frac{1}{2}\sin u\right]k^2$$

$$(8) \qquad + \frac{1}{4}\left[-\sin^2 u + \left(u - \frac{\sin u}{2}\right)^2\left(\frac{\csc u}{4}\right)(\csc^2 u + \cot^2 u)\right.$$

$$\left. + \frac{3}{4}\cot u \csc u\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right)\right]\frac{k^4}{2!} + \cdots.$$

$$sc(u, k^2) = \tan u - \frac{1}{4}\sec^2 u\left(u - \frac{\sin 2u}{2}\right)k^2$$

$$(9) \qquad - \frac{1}{4}\left[\frac{3}{8}(1 + \sec^2 u)\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right)\right.$$

$$\left. - \left(u - \frac{\sin 2u}{2}\right)\tan^2 u - \frac{1}{2}\left(u - \frac{\sin 2u}{2}\right)^2 \tan u \sec^2 u\right]\frac{k^4}{2!} - \cdots.$$

$$dc(u,k^2) = \sec u - \frac{1}{4}\tan u\left[2\sin u + \sec u\left(u - \frac{\sin 2u}{2}\right)\right]k^2$$

$$+\left[\frac{\sec u}{4}\left(u\sin 2u - \frac{\sin^2 2u}{2} - \sin^4 u\right) + \frac{1}{2}\tan^2 u\sin u\left(u - \frac{\sin 2u}{2}\right)\right. \tag{10}$$

$$\left. - \frac{3}{16}\sin u\sec^2 u\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right)\right.$$

$$\left. - \frac{1}{16}\sec u\left(u - \frac{\sin 2u}{2}\right)^2(\tan^2 u + \sec^2 u)\right]\frac{k^4}{2!} + \cdots.$$

$$nc(u,k^2)) = \sec u - \frac{\tan u\sec u}{4}\left(u - \frac{\sin 2u}{2}\right)k^2$$

$$+\left[\left(\frac{1}{16}\sec u + \frac{1}{8}\tan^2 u\sec u\right)\left(u - \frac{\sin 2u}{2}\right)^2 + \frac{1}{4}\tan^2 u\sin u\left(u - \frac{\sin 2u}{2}\right)\right. \tag{11}$$

$$\left. + \frac{3}{4}\sec^2 u\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right)\right]\frac{k^4}{2!} + \cdots.$$

$$nd(u,k^2) = 1 + \frac{\sin^2 u}{2}k^2 - \frac{1}{4}\left[u\sin 2u - \frac{\sin^2 2u}{2} - 3\sin^4 u\right]\frac{k^4}{2!} + \cdots. \tag{12}$$

TABLE 1

Computed Jacobian Elliptic Functions

u	$\sin^{-1}k$	*sn*					*cn*				
		k-Series		*u*-Series		Exact Value	*k*-Series		*u*-Series		Exact Value
		Computed Value	% Error	Computed Value	% Error		Computed Value	% Error	Computed Value	% Error	
.1745	5°	.17361	.02	.17361	.02	.17365	.98481	0	.98481	0	.98481
.1754	80°	.17365	0	.17365	0	.17365	.98481	0	.98654	.18	.98481
.7016	15°	.64277	0	.64294	.02	.64279	.76606	0	.76668	.08	.76604
.9401	50°	.76597	.01	.77886	1.67	.76604	.64282	0	.66704	3.77	.64279
1.1424	45°	.86587	.02	.90344	4.32	.86603	.50024	.05	.56037	12.07	.50000
1.3372	40°	.93964	.01	1.02184	8.74	.93969	.34250	.14	.45935	34.3	.34202
1.5738	5°	.99999	0	1.000821	.82	1.0000	0	0	.02496		0

u	$\sin^{-1}k$	*dn*				
		k-Series		*u*-Series		Exact Value
		Computed Value	% Error	Computed Value	% Error	
.1745	5°	.99986	0	.99988	0	.99988
.1754	80°	.98525	0	.98698	.17	.98527
.7016	15°	.98606	0	.98626	.02	.98606
.9401	50°	.80898	.09	.82828	2.29	.80972
1.1424	45°	.79147	.12	.83341	5.42	.79056
1.3372	40°	.80027	.41	.87352	9.61	.79697
1.5738	5°	.99620	0	.99837	.22	.99619

$$cd(u,k^2) = \cos u + \frac{\sin u}{4}\left(u + \frac{\sin 2u}{2}\right)k^2 + \left[-\frac{\cos u}{16}\left(u - \frac{\sin 2u}{2}\right)^2\right.$$

$$+ \frac{3}{16}\sin u\left(\frac{3}{2}u - \sin 2u + \frac{\sin 4u}{8}\right) \tag{13}$$

$$\left. - \frac{\cos u}{4}\left(u\sin 2u - \frac{\sin^2 2u}{2} - 3\sin^4 u\right)\right]\frac{k^4}{2!} + \cdots.$$

$$sd(u,k^2) = \sin u - \left[\frac{\cos u}{4}\left(u - \frac{\sin 2u}{2}\right) - \frac{\sin^3 u}{2}\right]k^2$$

$$- \left\{\frac{\cos u}{4}\left[\frac{3}{4}\left(\frac{3u}{2} - \sin 2u + \frac{\sin 4u}{8}\right) + \frac{\tan u}{4}\left(u - \frac{\sin 2u}{2}\right)^2\right]\right. \tag{14}$$

$$\left. + \frac{\sin u}{4}\left(u\sin 2u - \frac{\sin^2 2u}{2} - 3\sin^4 u\right)\right\}\frac{k^4}{2!} - \cdots.$$

Representative results computed for basic Jacobian functions using the k-series as well as the u-series are given in Table 1. The computations were carried to the first three terms for both types of series. The various k-u values, representing typical combinations in different ranges, were picked from the table of elliptic integrals [2], so that the "exact" values of the functions are known and that the relative accuracies of the two types of series may be compared. The percentage errors are computed to the nearest .01. It is seen that the k-series are nearly always more accurate than the corresponding u-series, particularly at larger u-values.

Transport Division
The Boeing Company
Renton, Washington

1. H. Bateman, *Higher Transcendental Functions*, v. II., McGraw-Hill Book Company, New York, 1953, p. 344.
2. E. Jahnke & F. E. Emde, *Table of Functions*, Fourth Edition, Dover, New York, 1945.

REVIEWS AND DESCRIPTIONS OF TABLES AND BOOKS

29 [C, D].—LEON KENNEDY, *Handbook of Trigonometric Functions—Introducing Doversines*, Iowa State University Press, Ames, Iowa 1961, 397 p., 23.5 cm. Price $6.50.

In his preface the author states that this new set of trigonometric tables has been arranged specifically to facilitate the solution of both plane and spherical triangles. The term "doversine" has been introduced as an abbreviation for "doubled versed sine."

The introduction to these tables contains a derivation of the law of haversines for spherical triangles and of the "law of doversines" for plane triangles. Also included therein are a table of algebraic signs for all the tabulated trigonometric functions for the four quadrants; a composite graph of the cosine, versine, haversine, and doversine; a compilation of fundamental trigonometric formulas and identities; and a presentation of formulas for solving plane and spherical triangles, including the navigational triangle.

The main table, occupying 360 pages, consists of six-figure natural and logarithmic values of the six standard trigonometric functions and of the versine and doversine. These data are conveniently displayed on facing pages for each sexagesimal degree to 180 degrees at increments of a minute in the argument. Natural values and their logarithms appear in adjacent columns, printed in black and blue, respectively. Differences are not tabulated.

Following this table, there is a table of six-place mantissas of the common logarithms of the integers between 1000 and 10,000.

The numerical values presented in these original tables were computed by the author on an IBM 650 system in the Ames Laboratory at Iowa State University. The doversines and their logarithms were previously computed on the ILLIAC, and these results were compared with corresponding data obtained in the later calculation.

The reviewer has compared nearly forty percent of the tabular entries appearing in the first quarter of the main table with the corresponding entries appearing in more elaborate tables such as those of J. Peters. The reliability of these trigonometric values and their logarithms may be inferred from the fact that such careful examination revealed just one typographical error (in csc 8°48′) and one terminal-digit error (in log sec 5°5′). On the other hand, the following five rounding errors were discovered by the reviewer in the table of logarithms of numbers: log 499, *for* 698100, *read* 698101; log 2443, *for* 387924, *read* 387923; log 8652, *for* 937116, *read* 937117; log 8854, *for* 947139, *read* 947140; and log 8884, *for* 948608, *read* 948609.

Correspondence with the author has disclosed that these last errors are attributable to a programming error, which led to the omission of one of the tests for retention of an adequate number of terms in the evaluation of the logarithms by power series. The publishers have recently informed the reviewer that they will make appropriate corrections in the future printings.

This handbook of trigonometric functions is unique among the tables that this

reviewer has examined, particularly because of its format, which is to be contrasted with the customary semiquadrantal arrangement. The tabulation of the versine and doversine clearly requires the extended range of the argument given in these tables. The author justifies the tabulation of the six standard trigonometric functions also over the first two quadrants on the basis of the resulting ease of application to the solution of triangles. The typography is generally excellent, and the arrangement of the tabular data convenient. This book should prove a useful addition to the literature of mathematical tables.

J. W. W.

30 [F].—R. KORTUM & G. MCNIEL, *A Table of Periodic Continued Fractions*, Lockheed Aircraft Corporation, Sunnyvale, California, 1961, xv + 1484 p., 29 cm.

This huge and interesting table contains, first, the *half-period* of the regular continued fraction for the $\sqrt{D}$ for each non-square natural number D less than 10,000. For example, since

$$\sqrt{13} = 3 + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{3 + \sqrt{13}},$$

under $D = 13$ are listed the *partial quotients*: 3, 1, 1. Again, since

$$\sqrt{19} = 4 + \frac{1}{2} + \frac{1}{1} + \frac{1}{3} + \frac{1}{1} + \frac{1}{2} + \frac{1}{4 + \sqrt{19}},$$

under $D = 19$ is the list: 4, 2, 1, 3.

Next, let

$$\sqrt{D} = q_0 + \frac{1}{q_1} + \frac{1}{q_2} + \cdots \quad \text{and} \quad x_i = q_i + \frac{1}{q_{i+1}} + \cdots$$

and $x_i = (\sqrt{D} + P_i)/Q_i$. Then Q_i, the *denominators of the complete quotients*, are listed in a row parallel to the q_i.

If p, the *period* of the continued fraction, is odd, as in $p = 5$ for $D = 13$, the table gives the smallest solution x, y of

$$x^2 - Dy^2 = -1.$$

If p is even, as in $p = 6$ for $D = 19$, the smallest solution is given of the so-called Pell equation:

$$x^2 - Dy^2 = +1.$$

The values of D for which p is odd are marked with an asterisk.

Finally, if $p^2/D > 1$, this ratio is given to 9 decimals. All of this was computed on an IBM 7090 in 36 minutes.

While in [1] the continued fractions for $D < 10{,}011$ have already been given, together with *both* sequences P_i and Q_i defined above, the range here of D for the solutions x, y, and for the ratio p^2/D, would appear to exceed that in any published table.

From a theoretical point of view the quantity p^2/D is of considerable interest. If we list those D where p^2/D attains a new maximum we obtain the following table:

D	p^2/D	D	p^2/D
3	1.333	631	3.651
7	2.286	919	3.917
43	2.326	1726	4.487
46	3.130	4846	4.768
211	3.204	7906	4.948
331	3.492		

Further, the authors have extended their computation to $D = 51{,}000$ and have listed all such D for which p^2/D exceeds 5. We may thus continue:

D	p^2/D	D	p^2/D
10651	5.141	19231	5.731
10774	5.257	32971	5.819
18379	5.641	48799	6.064

While these empirical data obviously suggest the possibility that

$$p = O(D \log D)^{1/2},$$

the authors refrain from such a suggestion and also from any reference to pertinent theoretical work.

Tenner's algorithm was used in the computation of the continued fractions. This requires two divisions and subtractions and one multiplication and addition in each cycle. An alternative algorithm is known that replaces one of these divisions by an addition. This latter computation would therefore be somewhat faster. But, alternatively, the redundancy in Tenner's algorithm (implicit in the fact that in one of these divisions the remainder is always zero) allows for a check at each cycle. But whether the authors utilized this check, or indeed made any check, is not indicated in their introduction.

There is a printing defect which could have been easily avoided. In a block of 10 decimal digits, if the high-order digit is a zero, it is printed as a blank. Thus, for $D = 801$, $x = 500002000001$ and $y = 17666702000$ are printed as

50 2000001
1 7666702000

This is because the binary to decimal integer conversion subroutine which was used *deliberately* causes such suppression of high-order zeros on the assumption that they will not be preceded by a significant digit. To circumvent such suppression one can add 10^{10} to each binary number before conversion to 10 decimal digits. This *fictitious* high-order 1 is an 11th digit, and will not be printed, since only the 10 low-order digits will be converted. Nonetheless, the routine is deceived, by its presence, into thinking that the high-order zeros are not now high order. A programmer who keeps on his toes can often outwit the makers of subroutines.

In a covering letter one of the authors indicates that this table is the second [2] in a series of fourteen, or more, number-theoretic tables. While a few of these duplicate, at least in part, some known tables, the latter are often on magnetic tape, or cost money, or are otherwise inaccessible. The entire proposed series will certainly be welcome to mathematicians working in number theory.

D. S.

1. Wilhelm Patz, *Tafel der regelmässigen Kettenbrüche*, Berlin Akademie-Verlag, 1955.
2. The first is *A Table of Quadratic Residues for all Primes less than 2550*. See RMT 35, *Math. Comp.*, v. 15, 1961, p. 200.

31 [I].—Herbert E. Salzer & Charles H. Richards, *Tables for Non-linear Interpolation*, 11 + 500 p., 29 cm., 1961. Deposited in the UMT file.

These extensive unpublished tables present to eight decimal places the values of the functions $A(x) = x(1 - x)/2$ and $B(x) = x(1 - x)(2 - x)/6$, corresponding to $x = 0(10^{-5})1$. This subinterval of the argument is ten times smaller than that occurring in any previous table of these functions.

These tables can be used for either direct or inverse interpolation, employing either advancing or central differences. In the introductory text are listed, with appropriate error bounds, the Gregory-Newton formula and Everett's formula, for direct quadratic and cubic interpolation, and formulas for both quadratic and cubic inverse interpolation, employing advancing differences and central differences. Examples of the use of these formulas are included.

The convenience of these tables is enhanced by their compact arrangement, which is achieved by tabulating $B(1 - x)$ next to $B(x)$. This juxtaposition, in conjunction with the relation $A(1 - x) = A(x)$, permits the argument x to range from 0 to 0.50000 on the left of the tables, while the complement $1 - x$ is shown on the right.

The authors note the identity $A(x) - B(x) \equiv B(1 - x)$, which can be used as a check on interpolated values of $A(x)$, $B(x)$ and $B(1 - x)$, and also as a method of obviating interpolation for $B(1 - x)$, following interpolation for $A(x)$ and $B(x)$.

Criteria for the need of these interpolation tables are stated explicitly, with reference to both advancing and central differences.

A valuable list of references to tables treating higher-order interpolation is included.

The authors add a precautionary note that this table is a preliminary print-out, not yet fully checked.

J. W. W.

32 [I, X].—George E. Forsythe & Wolfgang R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, John Wiley & Sons, Inc., New York, 1960, x + 444 p., 23 cm. Price $11.50

The solution of partial differential equations by finite-difference methods constitutes one of the key areas in numerical analysis which have undergone rapid progress during the last decade. These advances have been accelerated largely by the availability of high-speed calculators. As a result, the numerical solution of many types of partial differential equations has been made feasible. This is a development of major significance in applied mathematics.

The authors of this book have made an important contribution in this area, by assembling and presenting in one volume some of the best known techniques currently being used in the solution of partial differential equations by finite-difference methods. This, I am certain, has not been an easy task, owing to the fluid state of many of the theories in this field. For the same reason it is not possible, at the present state of flux, to write a book on this subject which will successfully withstand the test of time. The authors well recognize this point when they state in their introduction: "The literature on difference methods for partial differential equations is growing rapidly. It is widely scattered and differs greatly in viewpoint and character. A definitive presentation of this field will have to wait until the present period of intense development has come to at least a temporary halt."

The book contains an introductory chapter in addition to four major chapters, as follows:

Introduction to Partial Differential Equations and Computers

1. Hyperbolic Equations in Two Independent Variables
2. Parabolic Equations
3. Elliptic Equations
4. Initial-Value Problems in More than Two Independent Variables.

Topics covered within these chapters include the concept of stability, the method of characteristics, the numerical solution of problems involving shock waves, the theory of Lax and Richtmeyer, the solution of eigenvalue problems, the Young-Frankel theory of successive over-relaxation, and the method of Peaceman and Rachford.

The phenomenon of instability, which frequently arises to plague and invalidate many solutions of partial differential equations by finite-difference methods, is discussed in detail. However, this reviewer cannot, in good conscience, agree with the method of approach used in presenting this important and fundamental concept. The authors begin their discussion by stating: "Although the stability of difference equations has been amply discussed in the literature, one rarely meets precise definitions. The subject is therefore in need of further clarification." Subsequently the authors proceed to develop their own definition of stability, which in the opinion of the reviewer, is neither precise nor especially illuminating. The definition of stability is unnecessarily complicated by its tie-in with the concept of convergence and with the "cumulative departure," whose order of magnitude can, in the words of the authors, "rarely be exactly determined."

Notwithstanding any differences of opinion concerning the method of treatment of specific topics, the book is highly recommended as an authoritative and timely exposition of some of the most significant techniques currently available for the solution of partial differential equations by finite-difference methods.

H. P.

33 [I, X].—CHARLES JORDAN, *Calculus of Finite Differences*, Second Edition, Chelsea Publishing Co., New York, 1960, xxi + 652, 21 cm. Price $6.00.

This is a reprint of the second edition of the well-known book by Charles Jordan. The republication of this excellent text on the calculus of finite differences and its

availability at a reasonable cost should be welcomed by all students of numerical analysis.

H. P.

34 [K].—W. S. Connor & Shirley Young, *Fractional Factorial Designs for Experiments with Factors at Two and Three Levels*, NBS Applied Mathematics Series, No. 58, National Bureau of Standards, Washington, D. C., 1961, v + 65 p. Price $0.40.

This publication contains a collection of fractional factorial designs for experiments in which some factors are to be studied at two levels or conditions and others at three levels. It is the sequel to two other catalogs [1], [2] of designs in the National Bureau of Standards Applied Mathematics Series that contain, respectively, plans for m factors each at two levels, and plans for n factors each at three levels. This new document gives plans for the mixed series involving $(m + n)$ factors, where the m factors each at two levels and the n factors each at three levels are given for 39 combinations (2^m3^n) of positive integer values of m and n for which $5 \leqq m + n \leqq 10$.

For each design the following are given: number of effects estimated, number of treatment combinations employed in the design, fraction of complete factorial experimental plan, analysis, and construction.

The method of construction of designs is described in Section 2. Fractions are selected so that low-order interaction effects, including main effects, are aliased with each other as little as possible. Section 3 contains a description of the mathematical model, in which it is assumed that all interactions between three or more factors are nonexistent, and a procedure for estimating the parameters contained in the model. Section 4 contains a discussion of procedures to test hypotheses and to construct confidence intervals. A worked example of $2^3 3^2$ design is presented in Section 5.

Section 6 is devoted to six particular designs for which the interaction effects between factors at three levels are defined in a different manner from that of the other designs.

H. H. Ku

National Bureau of Standards
Washington, D. C.

1. National Bureau of Standards, *Fractional Factorial Experiment Designs for Factors at Two Levels*, NBS Applied Mathematics Series, No. 48, U. S. Gov. Printing Office, Washington, D. C. 1957.
2. W. S. Connor & Marvin Zelen, *Fractional Factorial Experiment Designs for Factors at Three Levels*, NBS Applied Mathematics Series, No. 54, U. S. Gov. Printing Office, Washington, D. C., 1959.

35 [K].—N. V. Smirnov, Editor, *Tables for the Distribution and Density Functions of t-Distribution ("Student's" Distribution)*, Pergamon Press Ltd., New York, 1961, 130 p., 28 cm. Price $12.50.

This book, which is Volume 16 in the Mathematical Tables Series of Pergamon Press, is a translation of the Russian work issued by the V. A. Steklov Mathematical Institute of the Academy of Sciences of the U. S. S. R. There are three main

tables of the (cumulative) distribution and density function of the ordinary (central) t-distribution presented to six decimal places for several ranges of the argument and parameter, followed by four auxiliary tables and one table of interpolation coefficients.

The density function of the t-distribution is

$$s_\nu(t) = K_\nu \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where

$$K_\nu = (\pi\nu)^{-1/2}\Gamma[\tfrac{1}{2}(\nu + 1)]/\Gamma(\tfrac{1}{2}\nu);$$

and the distribution function is

$$S_\nu(t) = \int_{-\infty}^{t} s(u)\, du.$$

Table I gives, in parallel columns, the values of $S_\nu(t)$ and $s_\nu(t)$ for $t = 0.00(0.01)3.00(0.02)4.50(0.05)6.50$, with $\nu = 1(1)12$, and for almost as many values of t with $\nu = 13(1)24$. Some of the printing is imperfect, one of the values, namely, $S_{21}(1.00)$ on page 6, having its fourth decimal place almost completely missing. Fortunately, this value of t is one of those repeated at the top of the page following, so that the missing digit, 6, is available.

Table II gives the distribution function $S_\nu(t)$ for large values of the argument, namely, $t = 6.5(1.0)9.0$, and $\nu = 1(1)10$.

Table III gives $S_\nu(t)$ for large numbers of degrees of freedom for $t = 0.00(0.01)2.50(0.02)3.50(0.05)5.00$, namely, for $\nu = 25(1)35$. The first two differences of the function are also given in convenient form alongside the functional values.

Table IV lists the values of $S_\nu(t)$, according to the values of $\xi = 1000/\nu$, for $t = 0.00(0.01)2.50(0.02)5.00$ (the limit 5.00 is omitted in the Contents and Introduction) for $\xi = 30(2)0$. The values of the ξ-entry are equivalent to the values $\nu = 33\frac{1}{3}$, $35\frac{5}{7}$, (etc.), 250, 500, ∞ (corresponding to the normal distribution $\Phi(t)$). This type of tabulation facilitates harmonic interpolation, which is necessary for large values of ν.

Table V gives values of the polynomials $C_i(t)$, $i = 1$ to 4, that occur in R. A. Fisher's formula for the distribution function in inverse powers of ν,

$$S_\nu(t) = \Phi(t) - \sum_{i=1}^{\infty} \frac{C_i(t)}{\nu^i},$$

for $t = 0.00(0.01)5.00$. These values, together with those of $C_3(t)$ that are given for a small number of values of t, were used in the calculation of Table III.

Table VI gives the values of K_ν and its common logarithm for $\nu = 1(1)24$. In addition, two other quantities, p_s and q_s, are tabulated for $s = 2(1)21$, for which no explanation is given in the Introduction or anywhere else in the volume. Apparently $p_s = 1/\sqrt{s}$, but the reviewer was unable to determine the meaning of q_s or the reason for tabulating these quantities.

Table VII gives the upper percentiles of the t-distribution corresponding to the probability levels $Q = .4, .25, .1, .05, .025, .01, .005, .0025, .001, .0005$, for $\nu = 1(1)30$ and various other values to 10,000, and ∞.

Table VIII gives, in concise form, the coefficient $B = k(1 - k)/4$ in Bessel's quadratic interpolation formula. By showing only the values of k for which the value of B changes in the third decimal place, the table is cut down to little more than one-tenth of the size otherwise necessary.

The introductory part of the volume includes brief sections on interpolation, with numerical examples, formulas for calculating and checking the tables, and a list of eight applications. It would have been very helpful to give examples of applications listed, along the lines of the Introduction to *Biometrika Tables for Statisticians*, Vol. I, for example, and to indicate needs for tabulations of such detail and accuracy. Several key references besides the one to Fisher in *Metron* would also have been helpful.

As already indicated, the quality of printing is less than perfect. The alignment of the columns is poor, making it unnecessarily difficult to read across the page to find the entry corresponding to a t argument. It might be pointed out that provision of a column for t at the right of each page as well as the one at the left would largely have alleviated this difficulty. Other minor shortcomings are lack of a heading at the top of each page to identify the table at a glance; omission of the subscript ν in $S(t)$ at the bottom of page 7 and inconsistency in showing the argument values, as in $t = 0(0.01)5.00$; writing $C(t)$ for $C_5(t)$ at the top of page 8, where also the reference to values of the parameter ν is irrelevant; omission of the prime in the derivative $S'(t)$ at the beginning of Section III on page 11; and erroneously writing the t-interval for Table II as 1.0 instead of 0.1 in both the table of contents and in the title on page 69. In Table III on page 75 the subscripts on the differences Δ should be written as exponents (the bar in $\bar{\Delta}^2$ means that the second differences are all negative).

In spite of such shortcomings, this volume represents the most detailed tabulation of Student's t-distribution available and, while it is not recommended for the general practitioner, will be indispensable to statisticians and others who require finely tabulated values for theoretical or other reasons.

J. LIEBLEIN

Applied Mathematics Laboratory
David Taylor Model Basin
Washington 7, D. C.

36 [L].—E. N. DEKANOSIDZE, *Tables of Lommel's Functions of Two Variables*, Pergamon Press, New York, 1960. 492 p. Price $20.00.

This is an English translation from the Russian. The original was reviewed in *MTAC*, v. 12, 1958, p. 239–240. The introduction contains several typographical errors, and these have been noted in a recent paper by J. Boersma "On the computation of Lommel's functions of two variables," in *Mathematics of Computation*, v. 16, 1962, p. 232.

Y. L. L.

37 [L].—V. N. FADDEYEVA & N. M. TERENT'EV, *Tables of Values of the Function* $w(z) = e^{-z^2}\left(1 + \frac{2i}{\sqrt{\pi}}\int_0^z e^{t^2}\,dt\right)$, *for Complex Argument*, translated by D. G. FRY, Pergamon Press, New York, 1961, 280 p., 26 cm. Price $15.00.

This is a translation from the Russian edition, which appeared in 1954. The present volume is essentially a tabulation of the error function in the complex plane, and, with $z = x + iy$, $w(z) = u(x, y) + iv(x, y)$, gives 6D values of u and v. In Table 1 the range is $0 \leqq x$ 3, $0 \leqq y \leqq 3$, with spacing in each variable of 0.02. In Table 2 the range is $3 \leqq x \leqq 5$, $0 \leqq y \leqq 3$, and $0 \leqq x \leqq 5$, $3 \leqq y \leqq 5$, with spacing in each variable of 0.1. A formula is presented so that the table is everywhere interpolable to an accuracy within two units in the last place. For interpolation about z_0, the formula uses the data at $z_0 \pm h$ and $z_0 \pm ih$, where h is the spacing. Let

$$\begin{aligned} 2\bar{\Delta}_x f(x, y) &= f(x + h, y) - f(x - h, y) \\ 2\epsilon &= (\bar{\Delta}_x u - \bar{\Delta}_y v) + i(\bar{\Delta}_y u + \bar{\Delta}_x v) \\ \Delta_x^2 f(x, y) &= f(x + h, y) - 2f(x, h) + f(x - h, y) \\ \tilde{\Delta}_x w &= \bar{\Delta}_x w - \epsilon. \end{aligned}$$

The interpolation formula reads

$$w(z) \sim w(z_0) + h\tilde{\Delta}_x w + \tfrac{1}{2}h^2\Delta_x^2 w,$$

and values of $\tilde{\Delta}_x u$, $\tilde{\Delta}_x v$, $\Delta_x^2 u$ and $\Delta_x^2 v$ are provided.

The foreword, written by V. A. Fok, enunciates applications of the tables to physical problems. Some properties of $w(z)$ are also given. The authors' introduction gives various representations of $w(z)$, including power series, asymptotic expansions, and continued fractions. The method of constructing and checking the table is discussed. To find values of $w(z)$ accurate to 6D outside the tabulated range, some approximations based on the continued fraction expansion are presented.

Other tables of the error function for complex argument have appeared since the original issue of the present tables. See, for example, *Math. Comp.*, v. 14, 1960, p. 83. In this table, as well as the one under review, z is in rectangular form. For tables of the error function with z in polar form, see *MTAC*, v. 7, 1953, p. 178 and *Math. Comp.*, v. 12, 1958, p. 304–305; v. 14, 1960, p. 84.

Y. L. L.

38 [L].—I. E. KIREEVA & K. A. KARPOV, *Tablitsy Funktsiĭ Vebera, (Tables of Weber Functions)*, Vol. 1, Vychislitel'nyi Tsentr, Akad. Nauk SSSR, Moscow, 1959, xxiv + 340 p., 27 cm. Price 37 rubles. [An English translation by Prasenjit Basu has been published in 1961 by Pergamon Press, New York. Price $20.00].

Weber's equation

$$y'' - \left(a + \frac{z^2}{4}\right)y = 0 \tag{1}$$

is satisfied when $-a = p + \frac{1}{2}$ by Whittaker's function $D_p(z)$. If $y(a, z)$ is a solution,

so also are $y(a, -z)$, $y(-a, iz)$, and $y(-a, -iz)$. The solutions for $a - 1, a, a + 1$ are connected by a linear recurrence relation.

If we replace a by $-ia$, and z by $\zeta e^{i\pi/4}$, the equation becomes

$$y'' - (\alpha - \tfrac{1}{4}\zeta^2)y = 0, \tag{2}$$

which also has real solutions for real α, ζ. This has no recurrence relation connecting solutions for α, $\alpha - 1$, $\alpha + 1$, though there is a complex relation connecting those for α, $\alpha - i$, $\alpha + i$.

The equation (1) is usually taken as basic in the complex variable theory of these differential equations.

There are several special cases worth mentioning.

(i) If $p = a - \frac{1}{2}$ is an integer $\geqq 0$, $D_p(z) = e^{-z^2/4}h_n(z)$,

where $$h_n(z) = (-1)^n e^{z^2/2}\frac{d^n}{dx^n}e^{-z^2/2}$$

is a Hermite polynomial.

(ii) If $p = -q - 1$ is an integer $\geqq 0$, a solution is $e^{z^2/4}h_n^*(z)$, in which $h_n^*(z) = (-i)^n h_n(iz)$, which is again a real polynominal.

(iii) The case of integral p also includes solutions which involve repeated integrals of $e^{\pm z^2/2}$.

(iv) When $a = -p - \frac{1}{2}$ is an integer, the solutions involve a finite series of Bessel functions of order $\dfrac{2k+1}{4}$ (k an integer) and argument $\dfrac{z^2}{4}$. For instance, when $a = 0$, and z is real, two solutions are

$$\sqrt{\frac{z}{2}}\,K_{1/4}\left(\frac{z^2}{4}\right) \quad \text{and} \quad \sqrt{\frac{z}{2}}\left\{I_{1/4}\left(\frac{z^2}{4}\right) + I_{-1/4}\left(\frac{z^2}{4}\right)\right\}.$$

All these are relevant to equation (1), with real a and z. With equation (2) the only real case of interest is when $\alpha = 0$, and ζ is real, in which case

$$\sqrt{\zeta}J_{\pm 1/4}\left(\frac{\zeta^2}{4}\right)$$

are both solutions.

These cases are all considered in [1], where there are also tables for real solutions to equation (2) with $\pm\alpha = 0(1)10$.

The tables now under review break new ground. They are concerned with the case where $p = -a - \frac{1}{2}$ is real, while $z = x(1 + i)$, with x real. The reality of p means that a real recurrence relation remains in existence connecting solutions for $p - 1$, p, $p + 1$. The differential equation now becomes

$$\frac{d^2y}{dx^2} - \{(2p + 1)i - x^2)\}y = 0$$

and the solutions are no longer real, except when $2p + 1 = 0$.

The tables give $D_p\{x(1 + i)\} = u_p(x) + iv_p(x)$ for $\pm x = 0(.01)5$, $p = 0(.1)2$, and $\pm x = 5(.01)10$, $p = 0(.05)2$, generally to 6 decimals, except that there are only 5 decimals when $|x| \geqq 7.5$ and $p \geqq 1.8$ simultaneously.

Since p is real, the functions remain elementary when $p = 0, 1, 2$. In fact,

$$u_0(x) = \cos\frac{x^2}{2} \qquad v_0(z) = -\sin\frac{x^2}{2}$$

$$u_1(x) = x\left(\cos\frac{x^2}{2} + \sin\frac{x^2}{2}\right) \qquad v_1(z) = x\left(\cos\frac{x^2}{2} - \sin\frac{x^2}{2}\right)$$

$$u_2(x) = -\cos\frac{x^2}{2} + 2x^2\sin\frac{x^2}{2} \qquad v_2(x) = \sin\frac{x^2}{2} + 2x^2\cos\frac{x^2}{2}$$

are given in the Introduction, formula (6′).

When $p = -\frac{1}{2}$, the equation (3) becomes a real equation and has real solutions; it is then a case of (2) with $\zeta = \sqrt{2}x$. However, $D_{-1/2}\{x(1+i)\} = u_{-1/2}(x) + iv_{-1/2}(x)$ is not real, and we find

$$\begin{aligned} u_{-1/2}(x) - iv_{-1/2}(x) &= 2^{-3/4}\sqrt{\pi x}J_{-1/4}\left(\frac{x^2}{2}\right) \\ &= \{2^{-1/4}\sqrt{\pi}/\Gamma(\tfrac{3}{4})\}U_1(x, 2) \\ &= 2^{-3/4}\{W(0, x\sqrt{2}) + W(0, -x\sqrt{2})\} \\ &= -2^{-5/4}\sqrt{\pi x}J_{1/4}\left(\frac{x^2}{2}\right) \\ &= -\{2^{-7/4}\sqrt{\pi}/\Gamma(\tfrac{5}{4})\}U_2(x, 2) \\ &= 2^{-5/4}\{W(0, x\sqrt{2}) - W(0, -x\sqrt{2})\} \end{aligned}$$

where $W(a, x)$ is the function tabulated in [1], while $U_1(x, 2)$ and $U_2(x, 2)$ are those tabulated by Smirnov [2, p. 121].

The case $p = -\frac{1}{2}$, or $a = 0$, is in many senses the "central" case; it gives the simplest Bessel function representation, and it is rather surprising that it is not tabulated in the work under review. True, it is obtainable from other published tables, but so also are the cases $p = \frac{1}{2}, \frac{3}{2}$ as we see below; yet these two values of p occur in the tables, although the cases $p = -\frac{3}{2}, -\frac{5}{2}$, of equal complexity and involving the same Bessel functions, do not.

For $p = \frac{1}{2}, \frac{3}{2}$, consider first the derivative $\dfrac{d}{dx}D_{-1/2}\{x(1+i)\}$;

we have

$$\begin{aligned} u'_{-1/2}(x) - v'_{-1/2}(x) &= -2^{-3/4}\sqrt{\pi x^3}J_{3/4}\left(\frac{x^2}{2}\right) \\ &= \{2^{-1/4}\sqrt{\pi}/\Gamma(\tfrac{3}{4})\}U'(x, 2) \\ &= 2^{-1/4}\{W'(0, x\sqrt{2}) - W'(0, -x\sqrt{2})\} \\ &= -2^{-5/4}\sqrt{\pi x^3}J_{-3/4}\left(\frac{x^2}{2}\right) \\ &= -\{2^{-7/4}\sqrt{\pi}/\Gamma(\tfrac{1}{4})\}U_2'(x, 2) \\ &= 2^{-3/4}\{W'(0, x\sqrt{2}) + W'(0, -x\sqrt{2})\} \end{aligned}$$

where the prime indicates differentiation with respect to the argument λx of the function, concerned, whether this be $+1$, -1, $+\sqrt{2}$, or $-\sqrt{2}$; for example, $W'(0, -x\sqrt{2}) = -\sqrt{2}\,\frac{d}{dx} W(0, -x\sqrt{2})$.

Then

$$u_{1/2} = \frac{x}{2}(u_{-1/2} - v_{-1/2}) - \frac{1}{2}(u'_{-1/2} + v'_{-1/2})$$

$$v_{1/2} = \frac{x}{2}(u_{-1/2} + v_{-1/2}) + \frac{1}{2}(u'_{-1/2} - v'_{-1/2})$$

and

$$u_{3/2} = x(u_{1/2} - v_{1/2}) - \tfrac{1}{2}u_{-1/2}$$

$$v_{3/2} = x(u_{1/2} + v_{1/2}) - \tfrac{1}{2}v_{-1/2}.$$

None of these representations as Bessel functions is mentioned in the work reviewed.

The introduction (the reviewer has relied here on the translation by Presenjit Basu for the English edition) is an interesting one, describing properties of the function relevant to its original computation and numerical use, and including five pages of graphs and diagrams illustrating the behavior of the functions. Interpolation is considered with respect to both x and p, and the use of the recurrence relation for $p < 0$ or $p > 2$ is explained. There is no mention in the Introduction of special applications of the particular solutions given of equation (3), and none is known to the reviewer, although there are many for the solutions of equations (1) and (2), where parabolic cylinder coordinates are appropriate.

The Introduction contains also two auxiliary tables. The first (p. xx, xxi) gives $a_r = \pm p(p-1)(p-2)\cdots(p-2r+1)/2^{4r}r!$, and $b_r = \pm p(p+1)\cdots(p+2r-1)/2^{4r}r!$ for $r = 0(1)3$; these are coefficients in asymptotic expansions. The other table gives $\Gamma\left(\frac{1}{2} - \frac{p}{2}\right) \Big/ 2^{1+(p/2)}\sqrt{\pi}$ and $\Gamma\left(-\frac{p}{2}\right) \Big/ 2^{1+(p/2)}\sqrt{2\pi}$. In each case ten-figure values are given for $p = .05(.05)1.95$, $p = 1$ excluded.

The tables are arranged so that each opening gives $u_p(x)$, $v_p(x)$, $u_p(-x)$, $v_p(-x)$ for four values of p, at interval 0.1 or 0.05, and 51 values of x at interval 0.01, with extra lines (up to three) at the foot to assist in interpolation. The openings are grouped into subtables, with p running from 0 to 2 (with $p = 2$ standing alone on the last page, facing the next title), and a single block of values of x covering half a unit; there are thus 20 such subtables.

The photographic printing of typescript is reasonably good and clear, although exhibiting the familiar variation in blackness that is rather too common.

J. C. P. Miller

The University Mathematical Laboratory
Cambridge

1. J. C. P. Miller, Editor, *Tables of Weber Parabolic Cylinder Functions*, Her Majesty's Stationery Office, London, 1955.

2. A. D. Smirnov, *Tablitsy Funktsiĭ Eĭri i Spetsial'nykh Vyrozhdennykh Gipergeometricheskikh Funktsiĭ ...* (*Tables of Airy Functions and of Special Confluent Hypergeometric Functions ...*) Izdatel'stvo Akad. Nauk SSSR, Moscow, 1955. Edition in English, Pergamon Press, New York, 1960. See *MTAC*, v. 12, 1958, p. 84–86.

39 [L].—A. V. Luk'ianov, I. B. Teplov & M. K. Akimova, *Tablitsy solnovykh Kulonovskikh funktsiĭ (funktsiĭ Uittekera) Tables of Coulomb Wave Functions (Whittaker Functions)*, Moscow, 1961, xi + 223 p., 26 cm. Price 1.63 rubles.

These tables are intended for persons working in the field of nuclear physics; they were compiled in the Mathematics Section of the Physics Department of the Moscow State University in a joint project with the Laboratory of Nuclear Reactions, Scientific Research Institute for Nuclear Physics, the Moscow State University. The tabular values were computed on a STRELA computer in the University Computation Center.

The authors give the basic solutions of Coulomb functions as the functions $F_l(\eta, \rho)$ and $G_l(\eta, \rho)$, where $\eta = \frac{Zze^2}{\hbar v}$, $\rho = kr$, l is the orbital moment of the particle with only integer values, k is the wave number at infinity, ze the charge of the particle, and Ze the charge of the nucleus in whose field the particle is moving, r is the spherical coordinate, and $\hbar$ is Planck's constant divided by 2π.

Compilation of these tables was justified on the grounds that these functions are not connected with other functions by simple relationships. Available tables of Coulomb functions are essentially intended for calculating Coulomb functions at the surface of nuclei. The tables in this book are intended not only for calculating Coulomb functions on the boundary of the nuclei, but also for calculating integrals of products of radial functions encountered in the theory of direct nuclear reactions.

The tables were calculated as follows. When $\rho = 1$ the values of $F_{14}(\eta, \rho)$, $F_{15}(\eta, \rho)$, $G_0(\eta, \rho)$, $G_1(\eta, \rho)$, and their first derivatives were found by means of the series

$$F_l(\eta, \rho) = \sum_{n=l+1}^{\infty} A_n(l, \eta)\rho^n,$$

$$G_l(\eta, \rho) = F_l(\eta, \rho)K(\eta) \log \rho + \sum_{n=-l}^{\infty} B_n(l, \eta)\rho^n.$$

The coefficients $A_n(l, \eta)$, $B_n(l, \eta)$, and $K(\eta)$ can be found by formulas given in *Tables of Coulomb Wave Functions*, Applied Mathematics Series Report 17, National Bureau of Standards, 1952, and in a paper entitled "Coulomb wave functions in repulsive fields," by F. L. Yost, J. A. Wheeler, and G. Breit, in *The Physical Review*, v. 49, 1936, p. 174–189. The values of these functions for the remaining values of ρ were obtained by integrating the differential equation

$$\frac{d^2y}{d\rho^2} + \left[1 - \frac{2\eta}{\rho} - \frac{l(l+1)}{\rho^2}\right] y = 0$$

by the Runge-Kutta method programmed for automatic selection of the interval and relative accuracy of 10^{-5}, and for the remaining values of l by the recursion formula

$$\frac{1}{l+1}\left[\eta^2 + (l+1)^2\right]^{1/2} U_{l+1} + \frac{1}{l}\left[\eta^2 + l^2\right]^{1/2} U_{l-1} = (2l+1)\left[\frac{\eta}{l(l+1)} + \frac{1}{\rho}\right] U_l$$

On the basis of an error check described in the Introduction, the authors believe that the errors in the tabulated values generally do not exceed 2 to 3 units in the

last decimal place, except where the function changes sign, in which case only the five tabulated decimal places are reliable. They also state that linear interpolation ensures 2 or 3 reliable decimal places, and fourth-order interpolation, 4 or 5 places. However, the accuracy of interpolation is considerably less in the region of ρ close to one, where it decreases with increasing l. In addition, 2 or 3 reliable decimal places can be obtained by linear interpolation in η, except for the region of values where $\eta > 1.5849$.

The tables for each of the functions $F_l(\eta, \rho)$, $G_l(\eta, \rho)$, and $G_l'(\eta, \rho)$ are divided into three groups corresponding to three intervals of changes in η. Part 1 (pages 2–33) gives five-place tables of values of the function $F_l(\eta, \rho)$ with $\rho = 1.0(0.2)20.0$, $\eta = 0$ (variable) 0.39811, $\log \eta = -\infty$, $-0.8(0.1) - 0.4$, and $l = 0(1)15$. Part 2 (pages 36–37) gives five-place tables of $F_l(\eta, \rho)$ for the same ranges of ρ and l, $\eta = 0.50119$ (variable) 1.5849, and $\log \eta = -0.3(0.1)0.2$. Part 3 (pages 70–101) gives five-place tables of $F_l(\eta, \rho)$ for the same ranges of ρ and l, $\eta = 1.9953$ (variable) 6.3096, and $\log \eta = 0.3(0.1)0.8$.

Values of the function $G_l(\eta, \rho)$ for the same ranges of ρ and l are given as follows: on pages 104–135 for $\eta = 0$ (variable) 0.39811, $\log \eta = -\infty$, $-0.8(0.1) - 0.4$; on pages 138–169 for $\eta = 0.5119$ (variable) 1.5849, $\log \eta = -0.3(0.1)0.2$; and on pages 172–203 for $\eta = 1.9953$ (variable) 6.306, $\log \eta = 0.3(0.1)0.8$. Values of the first derivative $G_l'(\eta, \rho)$ with respect to ρ, with $\rho = 1.0(0.2)20$, and $l = 0$ or 1, are given on pages 206–209, when $\eta = 0$ (variable) 0.39811, $\log \eta = -\infty$, $-0.8(0.1) - 0.4$; on pages 212–215, for $\eta = 0.50119$ (variable) 1.5849, $\log \eta = -0.3(0.1)0.2$; and on pages 218–221 for $\eta = 1.9953$ (variable) 6.3096, and $\log \eta = 0.3(0.1)0.8$.

References cited in the Introduction include only one Soviet source and 15 English-language sources. Among the latter are the National Bureau of Standards tables (Reports 17 and 3033), Japanese tables of Whittaker functions, and three articles by C. E. Fröberg.

The tables given in this book are arranged in a convenient manner, and the clear print adds to their attractiveness. This book should be useful to persons who who are now using the National Bureau of Standards tables.

E. J. OSTEN

David Taylor Model Basin
Applied Mathematics Laboratory
Washington 7, D. C.

40 [L].—M. E. SHERRY, *The Zeros and Maxima of the Airy Function and Its First Derivative to 25 Significant Figures*, Electronics Research Directorate, Air Force Cambridge Research Center, Bedford, Mass., April 1959.

Table 1 gives to 25S the first 50 values of a_s and $Ai'(a_s)$ for which $Ai(a_s) = 0$. Similarly, Table 2 gives a_s' and $Ai(a_s')$ for which $Ai'(a_s') = 0$. These extend the precision of tables recorded in a rather recent volume edited by F. W. J. Olver (see *Math. Comp.*, v. 15, 1961, p. 214–215). Tables 3 and 4, respectively, give to 25S the first 18 coefficients in the asymptotic series of arc tan $\{Ai(x)/Bi(x)\}$ and arc tan $\{Ai'(x)/Bi'(x)\}$.

Y. L. L.

41 [I, M].—M. V. Cerrillo & W. H. Kautz, *Properties and Tables of the Extended Airy-Hardy Integrals*, MIT Technical Report 144, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, November 15, 1951.

In a previous review (*Math. Comp.*, v. 15, 1961, p. 215) I stated that tables of the Airy function $Ai(z)$ were not available for z complex in polar form. This is not correct, and I am indebted to Dr. Nelson A. Logan of Lockheed Aircraft Corporation, Sunnyvale, California, for calling my attention to the tables of Cerrillo and Kautz, here described.

Let

$$Ah_{1,3}(B) = 3^{-1/3}Ai(-3^{-1/3}B),$$

$$Ah_{2,3}(B) = \tfrac{1}{2}\,3^{-1/3}[Ai(-3^{-1/3}B) - iBi(-3^{-1/3}B)],$$

$$Ah_{3,3}(B) = \tfrac{1}{2}\,3^{-1/3}[Ai(-3^{-1/3}B) + iBi(-3^{-1/3}B)],$$

where $Ai(z)$ and $Bi(z)$ are the usual notations for the Airy integrals. Let $B = |B|e^{i\beta}$. This report gives the real and imaginary parts of $Ah_{1,3}(B)$, $Ah_{2,3}(B)$, and $Ah_{3,3}(B)$ to 7D for $|B| = 0(0.2)4$, $\beta = 0(7.5°)180°$. The functions $|Ah_{1,3}(B)|$ and arg $Ah_{1,3}(B)$ (in radians) are also tabulated to 7D for the same range, and graphs of these functions are also provided. The headings in each table should read $|B|$ for B and $|Ah_{1,3}(B)|$ for $Ah_{1,3}(B)$. Also tabulated to 7D are the values of the first 31 coefficients in the power series of $Ah_{1,3}(B)$ and the first 20 zeros of the latter for B real.

Y. L. L.

42 [L, M].—Herbert Bristol Dwight, *Tables of Integrals and Other Mathematical Data*, 4/e., The Macmillan Company, New York, 22 cm. x + 336 p. Price $3.50.

Reviews of the first two editions of these tables, published in 1943 and 1947 respectively, have appeared in *MTAC* (v. 1, p. 190–191; v. 2, p. 346).

A third edition, published in 1957, was enlarged through the addition of formulas relating to determinants, a more extensive list of derivatives of inverse trigonometric functions, a supplementary table of values of the exponential functions, and tables of natural values of the trigonometric functions (to 5D or 5S) corresponding to angles expressed degrees and hundredths.

The fourth and latest edition represents a further significant enlargement, the principal amplification being in the tabulation of definite integrals. This section now occupies 42 pages as contrasted with 11 pages in the third edition and eight pages in the first edition. The principal source of this information is cited as *Nouvelles Tables d'Intégrales Définies* by Bierens de Haan, Leyden, 1867, now readily available through republication in 1957 by Hafner Publishing Company in New York. The section devoted to elliptic functions has been extended by the inclusion of additional formulas concerning indefinite integrals expressible in terms of elliptic integrals.

Eleven references have been added to the list of 65 appearing in the third edition. A few minor changes have been made in the numerical tables; the most conspicuous

is the increase in precision of Table 1060 (Some Numerical Constants), so that all the entries appear now to 10 decimal places.

Errata pointed out in the earlier reviews have been corrected. However, the correction of $K(87.^{\circ}6)$ in Table 1040 was not followed by appropriate changes in the column of first differences. The reviewer has compared the entries in Table 1050 with the corresponding data in the tables of Lowell [1], and thereby has detected 72 last-figure errors in Dwight's values of the Kelvin functions of zeroth order and of their first derivatives.

This useful new edition of Professor Dwight's popular tables of integrals constitutes a valuable contribution to the increasing store of such mathematical literature.

J. W. W.

1. HERMAN H. LOWELL, *Tables of the Bessel-Kelvin Functions Ber, Bei, Ker, Kei, and their Derivatives for the Argument Range 0(0.01)107.50*, Technical Report R-32, National Aeronautics and Space Administration, Washington, D.C., 1959. (See Review 9, *Math. Comp.* v. 14, 1960, p. 81.)

43 [M].—W. F. HUGHES & F. T. DODGE, *A Table of J Integrals of Hydrodynamic Lubrication Theory*. Manuscript deposited in UMT file.

This unpublished table of the numerical values, mainly to five significant figures, of the integrals $J_n = \int_0^{\theta} (1 - \epsilon \cos \theta)^{-n} \, d\theta$, corresponding to $n = 1, 2, 3$, $\epsilon = 0.1(0.1)0.9$, and $\theta = 0^{\circ}(5^{\circ})360^{\circ}$, was prepared on an electronic digital computer by members of the Mechanical Engineering Department of the Carnegie Institute of Technology.

In the prefatory text the authors state that these integrals occur in the theory of the hydrodynamic lubrication of the journal bearings. The film thickness h is approximated by the formula $h = c(1 - \epsilon \cos \theta)$, in terms of the angular coordinate θ, the radial clearance c, and the ratio ϵ of the eccentricity of the journal to the radial clearance. Values of the ratio h/c to four decimal places are included in the table.

J. W. W.

44 [M].—G. PETIT BOIS, *Tables of Indefinite Integrals*, Dover Publications, Inc., New York, 1961, xiv + 151 p. 24 cm. Price $1.65.

This is a new printing, in an inexpensive paperback edition, of the original *Table d'Intégrales Indéfinies* published by Gauthier-Villars in Paris in 1906, and at the same time by Teubner in Leipzig under the title *Tafeln unbestimmte Integrale*.

This unabridged English translation contains 2544 indefinite integrals, systematically arranged according to integrands, as outlined in the table of contents. A preface lists the principal source books and tables. This is followed by an explanatory section devoted to notation and by a section listing 49 "transformations of integral expressions," that is, pairs of expressions possessing the same derivative.

With few exceptions, the indefinite integrals listed here involve elementary functions. Several integrals are shown to depend upon the evaluation of such functions as the sine and cosine integrals, although these are not identified as such. Examples of such higher transcendental functions, which are left in the form of

integrals without comments, appear on pages 118, 122, 123, 140, and 150. Furthermore, except for a footnote on page 62, elliptic integrals are nowhere referred to in these tables.

In the foreward to their extensive collection of indefinite integrals Gröbner and Hofreiter [1] refer to these tables as one of the sources for their material. Nevertheless, the tables of Petit Bois are not comparable to these recent German tables, nor indeed to the recently enlarged compilation of Dwight [2], principally because these last two tables cover a much larger spectrum of classes of functions.

The reviewer noted only one serious error of commission, namely, on page 150 appears an evaluation of the indefinite integral of $\log (a + \cos x)$ which is manifestly incorrect.

In conclusion, this reviewer considers this compilation to supplement to some degree the information presented in several more recent tables, such as those cited; nevertheless, it cannot replace them in general use.

J. W. W.

1. W. Gröbner & N. Hofreiter, *Integraltafel. Erster Teil: Unbestimmte Integrale*. Vienna and Innsbruck, Springer-Verlag, 1949.
2. H. B. Dwight, *Tables of Integrals and other Mathematical Data*, fourth edition. The Macmillan Co., New York, 1961.

45 [P, Z].—Robert S. Ledley, *Digital Computer and Control Engineering*, McGraw-Hill Book Co., New York, 1960, xvii + 835 p., 24 cm. Price $14.50.

The author's purpose in writing this book, as stated in the Preface, is "to fill the need for a comprehensive, elementary engineering textbook in the large and still growing field of digital computers and controls." Without doubt, it is comprehensive; more than enough is included in over 800 pages for a year's course or for several years of graduate work if the Additional Topics are included. The author, is justifiably proud of the 750 exercises scattered throughout the text. The exercises form a framework to hold the book together and permit some of the author's objectives to be achieved, while the additional topics provide opportunities for rich learning experiences for honors or graduate students, as well as insights for the exceptional student seeking more than grades and credits.

The book is divided into five parts, each one reasonably independent and self-contained.

Part 1, entitled Introduction to Digital Programming Systems, serves to introduce digital systems and to stimulate interest in them through examples and the theory of their applications. It may be that an unwary and not too disciplined student, not familiar with the field at all, may be frightened away by the rather sophisticated examples. Chapter I does show, however, that digital computers and controls are tools—the means to realize highly complex programs. There is not, in the first chapter, on the other hand, any hint of what these tools are or what they actually do, or how. This is probably a good technique, for now curiosity alone should lead the student to Chapter 2 to discover the secrets of such remarkably versatile hardware. This chapter, however, will probably have to be read twice to be understood by most neophytes. As all good computer people do, Professor Ledley personifies the hardware; for example, "the computer is told"; control "interprets" and "tells ... the arithmetic unit what to do," etc. Without a previous knowledge

of the physical meanings and the allegorical nature of these personifications, a good number of questions may arise. Here, as throughout the text, the exercises do much to dispel such problems. The exercises are vital to the complete understanding of the material in this text, because of its often descriptive nature. Where the descriptions or analogies are weak, such as the use of the terms "read-in" and "read-out," "memorizing information," and the sentence "memory boxes are analogous to mail boxes," one must hope that the instructor can salvage the situation. Getting into an unfamiliar discipline is first of all a matter of mastering a new vocabulary; here the student will stumble a while before finding this key to new knowledge.

The input, output, and memory section of Chapter 2 is descriptive. The approach is to cite an application, show a system that can perform it, and then discuss the units making up such a system. This is logical, but the engineer will be given electrical specifications to meet, not end applications. All of the picture should be presented to him, but the greatest emphasis should be placed on where he will go—on the design of electronic circuits. This does not happen here. (An unusual feature in this book is that the history of computers is relegated to an additional topic of Chapter 2, accompanied by an interesting evolutionary tree showing computer development.) This chapter is rich in photographs, but the block diagrams are likely to carry far more meaning to the student.

In Chapter 3, after much assurance to the student that all this is really necessary, the book gets into some serious work. This chapter, concerned with coding and programming, is highly recommended. It contains numerous excellent presentations on number systems, binary arithmetic, coding, flow charting, symbolic code, multiple-address systems, decimal-coded-binary systems, more excellent exercises, and extremely interesting additional topics.

In Chapters 4 and 5, programming fundamentals and advanced programming, respectively, are both presented with clarity, and again are highly recommended, especially for prospective programmers. While it may be that so much attention to programming is discouraging to the engineer who is primarily concerned with hardware, the need for ingenuity, precision, and economy in coding is well demonstrated. This may serve to bring into sharper focus the similarity between the user's or programmer's requirements and the engineer's own disciplines.

Part 1, then, despite an uncertain start, should leave the student with few doubts about the why of computers.

Part 2, the Functional Approach to Systems Design, approaches the subject from a mathematical rather than a purely descriptive level. This Part will prove to be a trial for most engineering students, for still there is no hint of hardware. However, as another contribution to the building of a solid foundation, it, too, has great value.

Chapter 6, Fundamentals of Numerical Analysis, presents several discussions of how problems may be prepared for systematic solution by computers. The section on accuracy and errors should dispel naive notions of the infallibility of computers.

Chapter 7, entitled Searching, Sorting, Ordering, and Codifying, is another very good introductory chapter, this time to data processing. Section 7-7, Codifying, concerns itself with error correction and superimposition. The discussion of self-

correcting and redundant codifications should prove of value in the engineer's career.

In Chapter 8 are found discussions of several special-purpose digital-computer system designs. These are all good, and again should broaden the student's outlook. With the nearly 300 pages so far devoted to introductions and elaborations, he may, however, have given up finding out just what makes these machines operate.

In Chapter 9, the system design of the Pedagac ("pedagogic automatic computer") is presented. None of the reasoning leading to this particular configuration is shown, leaving the student to accept it as a *fait accompli*. The appended Additional Topics present some fine challenges to student programmers, namely, the writing of an automonitor, a library of subroutines, and an ALGOL for Pedagac.

Part 3 deals with the Foundation for the Logical Design of Digital Circuitry. Chapter 10 provides a good, practical introduction to Boolean Algebra—a necessity for digital design. Propositional Calculus is brought in by way of background material, and is followed by a rapid, clear, and concise development of Boolean Algebra. The all-important parallel between computer logic and Boolean logic is firmly established, and a fine set of Additional Topics completes this chapter.

Chapters 11 and 12 deal with numerical representations of Boolean expressions and with many elegant computational methods of simplification and logical design, often the result of original work by the author and his associates. These are likely to be of much value to the serious computer designer. Chapter 13, entitled Boolean Matrix Equations and Fundamental Formulas, presents more simplification and computational methods. It is in this part of the book that Dr. Ledley reaches a peak of accomplishment. In Sections 11-9, 11-10, and in all of Chapters 12 and 13, it is design that is emphasized, not analysis or description. It was R. W. Johnson in the August 1961 edition of the IRE Proceedings who asked "Are electronics engineers educated?"—stressing how much analysis is taught and how little design, which is the life blood of engineering. In Part 3 of this book, Dr. Ledley teaches design. He achieves a clarity of presentation, never resorting to "mathmanship" or other "impressive confusion." The advantages of an interdisciplinary approach, at the elementary level (for example, algebra taught by a Professor of Engineering) rather than the highly sophisticated specialist trying to "write down" to students, are amply illustrated. In fact, the only quarrel is with the nomenclature, the use of the term "circuit" design when actually "logical" or "symbolic" design is meant. Chapter 14, Applications of Matrix Equations in Circuit Design, concerns itself with applying the methods and principles just learned to solving specific problems. (This chapter and its Additional Topics are unlikely material for engineering students, but are certain to have high appeal to logicians and mathematicians.)

Part 4, paradoxically named Logical Design of Digital-Computer Circuitry, reveals to the student how all the Boolean tricks he has learned may be used to group components logically to make them compute. Chapter 15 presents the logical realization of serial arithmetic. It is in this chapter that the long-postponed revelation—how does a computer actually work—is made; that is, what is it that a computer does when it computes? How is computation mechanized? The presentation is excellent.

Having exposed the "heart" of computers, the author follows through with

some extremely important and interesting elaborations in Chapter 16, Parallel and Rapid Arithmetic Operations. The student will again be challenged. Equally important, but not as obvious, is the control logic of a computer, lucidly presented in Chapter 17. Chapter 18 is unfortunate in that the packaging and logical design of the Pedagac are presented in such rapid order that the "thread of continuity" aimed for by the author is violently contradicted. A complete logic for the computer is displayed, rather than being built-up or designed in a step-by-step fashion. It appears as though after the foundation was built painstakingly and in great detail, the finished product is suddenly presented, fully built.

Part 5, Electronic Design of Digital Circuits, is the only part of the book dealing with electronics, and in a somewhat superficial manner, at that. Chapter 19, which discusses Problems and Limitations in Electronic Realization, includes a good presentation of digital gating systems. Static and dynamic storage and parametric amplifier methods are described clearly. No circuits are designed in this chapter, but the important topics of clock phasing, synchronization, and circuit reliability are discussed. An interesting Boolean-probabilistic method of logical design for increased reliability is developed. The preponderant use of semiconductors in today's equipment is acknowledged by a full discussion of their physical theory in Chapter 20, Semiconductor Elements in Digital-circuit Design. Both large- and small-signal equivalent circuits for transistors are shown, and some circuit analysis is carried out. Diode gates, transistor gates, and current switching are described. A typical transistor flip-flop circuit is analyzed, and finally, too, the physical theory and parametron analogies of tunnel diodes are discussed.

Continuing at the introductory level, Chapter 21 presents a review of magnetism, and a discussion of square-hysteresis-loop cores. The sections on magnetic amplifiers, magnetic core logic, and multi-aperture cores are excellent, and will prove useful in an engineer's later work. Again, an unusual and highly desirable feature here is the rather thorough discussion of recent developments such as microwave and cryotron computer components, including exercises in their use as logical elements.

Memory and input-output methods are described in Chapter 22, which includes a section on magnetic film memories. The discussions of coincident current techniques, the several methods of recording and reading from magnetic tape and drums, analog-to-digital and digital-to-analog conversions, visual displays, page scanners and line printers, in the 30-odd pages of this chapter, provide a wealth of introductory material for an interested student. The discussion of the sampling theorem for analog-to-digital conversion must be cited as also highly desirable. The exercises for this chapter concentrate on logic and programming, and provide no further electronic circuit design experience.

Chapter 23 completes a major work of this book, the electronic design of the Pedagac. Unfortunately, this device for providing continuity serves rather to disrupt the continuity. Three widely scattered chapters present descriptions of the complete machine, that is, the system, the detailed logic, and the hardware. The final design is displayed in completed form rather than being built up step by step, and although it serves as good illustrative material for developing important concepts clearly and concretely, an opportunity for making the book truly continuous is thereby narrowly missed. This, however, is the lesser of two criticisms; the other

is the extraordinarily large proportion of logic, mathematical theory, and purely descriptive matter to the material dealing peripherally with electronic design problems and their solutions.

Such might include the many input-output synchronization problems, the numerous problems of cross-talk and pick-up, the signal-to-noise ratio of input devices, the power requirements, the transmission of pulses over relatively long lines, failure localization, ground loops, and so forth, which every computer design engineer must face. Special circuitry, such as sense amplifiers, clock pulse generators, and magnetic-core switching-current pulse-generators are not mentioned, and, of course, may be found elsewhere. The point is, however, that these are the problems likely to haunt the engineer long after his Boolean expressions have been mechanized on paper. That is why this is a difficult book to evaluate. It is at once an extremely broad survey of the "state of the art," reflecting a familiarity with a good deal of the current literature and at the same time a thorough exposition of many of the basic principles and practices underlying the digital computer technology. As a textbook, it should prove valuable, for it is written to and for students, with illustrations and clarifications used profusely. This volume may also prove to be of value to those whose experience in the computer field, while long, may have been rather narrow—technicians, circuit designers, programmers, etc. One may question, however, whether this admirable work actually belongs in the publisher's Electrical and Electronic Engineering Series. For all its strong points, and they are many, the book's weakest area is precisely that of electronic or electrical engineering. It contains numerous excellent presentations of basic principles, and contributes greatly to the diffusion of knowledge in logical design methodology. These are valuable for the engineer's background and understanding, but are more likely to be used by mathematicians or logical designers. As a text in electrical and electronic engineering this otherwise remarkable book can hardly be placed alongside the books of Millman & Traub, Millman & Seely, Terman, and Skilling. It is a book that can stand on its own merits, and needs not to lean on an Electrical and Electronic Engineering Series, achieved by a title with the word Engineering in it. An alternative name for the book, such as "Elements of Digital Computer Programming and Logical Design," would not raise false hopes nor be misleading. This is perhaps a criticism of the publisher and editors and not properly of the book, but must be made, nevertheless.

HERBERT ERNST

Applied Mathematics Laboratory
David Taylor Model Basin
Washington 7, D. C.

46 [Q].—DIRK BROUWER & GERALD M. CLEMENCE, *Methods of Celestial Mechanics*, Academic Press, New York, 1961, xii + 598 p., 24 cm. Price $15.50.

The title of this book appears to represent a compromise with an aspiration that must have had its origin more than a quarter of a century ago. In 1943 this reviewer overheard Bart Bok say, "I told Dirk the best thing he can do for Astronomy is to finish his book." Nearly ten years ago Brouwer was joined by the second author, with the same objective still in view. Both authors have been close collaborators in numerous astronomical endeavors since the end of World

War II. Both have also been amongst the most active research workers in celestial mechanics. This book is unequivocably the finest volume to appear in its field in the English language in the present century. The compromise referred to above is simply that it is not an exhaustive treatise. Nor is it a volume which could have appeared two decades ago. It sets forth the combined wisdom of the authors, based on their extensive experience up to the present, and includes some material (dealing with artificial satellites) not in existence two or three years ago.

The early chapters cover the usual introductory material: elliptic motion, expansions, and attraction of finite bodies. Then follow a few "practical" chapters on finite differences, numerical integration, aberration, precession, nutation, least squares, etc. The meat of the volume is in the chapters entitled General Integrals, Variation of Arbitrary Constants, Lunar Theory, Perturbations of the Coordinates, Hansen's Method, Disturbing Function, Secular Perturbations, and Canonical Variables. While the material is the same, there is nothing of the stereotyped presentation of classical treatises. It is here that the authors have given their own distinctive flavor to the work. Many features could be cited; for example, a deliberate effort to present Hansen's method in its most favorable light, and a correction method for deriving trigonometric series for the negative powers of the distance.

Whether this will prove to be a good textbook in an advanced graduate course is not easy to say without having tried it. That it will prove to be a valuable reference volume for a wide variety of workers both inside and outside this field of specialization is indubitable. For the long-term good of Celestial Mechanics this reviewer is of the opinion that an advanced text is needed which presents the material in the terminology that is familiar to present-day graduates in mathematics and physics; that is, vectors, matrices, gradients, dyadics, tensors, etc. But this is a different objective from that which the authors have set for themselves.

The volume has been meticulously prepared, both by the authors and the publisher, and competently proofread by G. Hori. If errors exist, it would be a shame, but reviewers delight in spotting what they can. To set the record straight, we note that Cowell's method (p. 186) had the incentive of its origin in the discovery of the eighth satellite of Jupiter (*Monthly Notices of the Roy. Astr. Soc.*, v. 68, p. 576). The calculations for Halley's comet came later. But such items cannot detract from the permanent value of this monumental volume.

PAUL HERGET

Cincinnati Observatory
Cincinnati, Ohio

47 [S].—A. N. ZAIDEL', V. K. PROKOF'EV & S. M. RAISKII, *Tables of Spectrum Lines*, Pergamon Press Ltd., Oxford, 1961, xliii + 554 p., 24 cm. Price $14.00.

These tables, compiled largely from the M. I. T. Wavelength Tables, were first published in Moscow in 1952. In 1955 the tables were reprinted in Berlin as an International Edition, with introductory text in German, English, and French. The new edition appears to be identical with the Berlin edition except for rearrangement of the introductory sections.

The tables are in three parts. Part I lists the spectral wavelengths of 32 of the more common elements in order of decreasing wavelength between 8000 and 2000A. Intensities in arc, spark, and discharge tube are given, and air lines are included.

Part II lists the stronger lines and intensities of 93 elements. Excitation energies in electron volts are given for about 75 per cent of the lines. In Part III are presented several short auxiliary tables, mainly of physical properties and sensitive spectral lines, which are designed especially to aid in spectrochemical analysis. These include ionization potentials, atomic and molecular weights of the elements and their oxides, with their melting and boiling points, a table of the elements in sequence of their appearance in the spectrum of the carbon arc, a short list of sensitive lines of the elements in order of wavelengths and by elements, a table of strong lines between 2000 and 1800A, a table of iron lines suitable for intensity standards, and spectra of the hydrogen molecule and of deuterium.

A major feature of this book lies in the selection and arrangement that was made of the more important data on wavelengths and intensities. Many weaker lines, especially in the spectra of the less common elements, were omitted. The inclusion of excitation energies in Part II provides the first extensive compilation of such data in convenient form. These features and the useful auxiliary tables should be helpful particularly to the spectrochemical analyst. However, some caution should be observed in using the tables, since errors that were present in the M. I. T. tables may be found in this book; for example, 2592.627 Cu I, intensity 1000, is in error and should appear as 2392.627. The translations between languages may result in misspelling; for example, on page xx, Muir should be Moore (C. E.). However, these minor errors do not detract seriously from the general usefulness of the book.

BOURDON F. SCRIBNER

National Bureau of Standards
Washington, D. C.

48 [S, V].—L. S. BARK, P. P. GANSON & N. A. MEISTER, *Tablitsy skorosti zouka v morskoĭ vode* (*Tables of the Speed of Sound in Sea Water*), Moscow, 1961, xiii + 182 p. + inserts, 26 cm. Price 1.73 rubles.

These tables of the speed of sound in sea water were computed in 1960 on a Strela-3 computer and a T-5 tabulator in the Computation Center at the request of the Institute of Oceanology supported by the Institute of Acoustics (all of the Academy of Sciences USSR).

Since the authors believe that previous tables, published by the British Admiralty [1], were based on formulas less accurate than Del Grosso's formula [2], they have based their tables on the Del Grosso formula

$$v = 1448.6 + 4.618t - 0.0523t^2 + 0.00023t^3 + 1.25(S - 35) - 0.011(S - 35)t + 0.0027 \times 10^{-5}(S - 35)t^4 - 2 \times 10^{-7}(S - 35)^4(1 + 0.577t - 0.0072t^2),$$

where t is the temperature in degrees Centigrade and S is the salinity in parts per 1000.

The tables consist of two parts, supplemented by appendices consisting of eight nomograms of the speed of sound and by tables of corrections for depth. These appendices are inserted in a cover pocket. The tables are divided into two parts because the Del Grosso formula yields a guaranteed accuracy of 0.2 m./sec. only when the salinity is more than 19 parts per 1000. The first part of the tables (pages 1–36) gives values of the speed of sound for a range of temperatures from

$-2°$ to 33°C and salinity from 0 to 20 parts per 1000. This ensures precision in the determination of the speed of sound in water, within this range of salinity, of 0.6 m/sec without interpolation.

The second part of the tables (pages 37–180) covers the same temperature range and salinity ranging from 20 to 40 parts per 1000. The steps are, respectively, 0.05°C and 0.2 part per 1000; thus, the precision of determination of the speed of sound without interpolation is here 0.3 m/sec.

The nomograms are divided into two groups according to salinity, the first group containing Nomograms 1 to 3 for salinity ranging from 0 to 20 parts per 1000, and the second group containing Nomograms 4 to 8 for salinity from 20 to 40 parts per 1000. The range of temperature represented in these nomograms is $-2°$ to 30° Centigrade.

Since the values given in the tables and nomograms are values for zero depth in sea water, it is necessary to introduce the correction, Δv_p, for depth. These values are given for depths ranging from 0 to 10,990 meters. They are always positive, and are considered constant at 0.0161 m./sec. for depths up to 100 meters. The tables of corrections (pages x to xii, also one insert page) are based on the contents of W. D. Wilson's paper entitled "Speed of sound in sea water as a function of temperature, pressure and salinity," in the *Journal of the Acoustical Society of America*, v. 32, 1960, p. 641–644. Another basic assumption used in compiling the correction table is that a depth of 10 meters corresponds to a hydrostatic pressure of 1 kg/cm^2.

The references include three British, four Soviet, and five American authors.

The format of the tables is clear and convenient. Since only a few table and column headings are involved and the format of the tables is clear, these tables could be of value to persons who make use of similar publications in the English and German languages.

E. J. Osten

Applied Mathematics Laboratory
David Taylor Model Basin
Washington 7, D. C.

1. *Tables of the velocity of sound in pure and sea water for use in echo sounding and sounding ranging*, Hydrographic Department of the Admiralty, London, 1927; also D. J. Mathews, *Tables of the velocity of sound in pure water and sea water*, 2nd ed., 1939; 3rd ed., 1944, Hydrographic Department of the Admiralty, London.

2. V. A. Del Grosso, *The Velocity of Sound in Sea Water of Zero Depth*, Report N 4002, Naval Research Laboratory, Washington, 1952.

49 [S, X].—Rudolph E. Langer, Editor, *Partial Differential Equations and Continuum Mechanics*, The University of Wisconsin Press, Madison, 1961, xv + 397 p., 24 cm. Price $5.00.

This volume prints in full the invited papers and in abstract the contributed papers at a symposium in 1960. As to be expected in such a case, the quality of the papers is uneven. As a whole, the papers on pure analysis represent the state of the art in their field much better than those on continuum mechanics, which are also far fewer in number. The typing is as attractive as typing can be, and there are fewer misprints than usual, but the reviewer fails to see why the Department of the Army of the vaunted richest country in the world cannot afford ordinary printing as used by our poorer neighbors, or what purpose is served by publication of

these papers here rather than through the normal channels of selection by the numerous journals.

Among the many excellent papers on the general theory of partial differential equations should be mentioned those of Leray, Moser, Hörmander, and Morrey. An interesting new type of problem is set up and solved by B. Frank Jones; in this problem, a coefficient function in the partial differential equation is to be determined by initial and boundary data that would lead to an overdetermined problem if the coefficient were fixed. This type of problem obviously has great use in applications to complicated systems whose constitution is only partly known; results of this kind make possible inferences about the internal nature of the system from use of observations on the boundaries.

Three papers concern directly techniques or error estimates in numerical calculation. Jim Douglas, Jr. sets up an alternating direction method for iterative solution of Laplace's equation in any number of dimensions. From the error estimates he concludes that this method is the most efficient known for Laplace's equation over a rectangular parallelepiped. The brothers Nitsche obtain error estimates for the computation of certain integrals in whose integrands occur solutions of elliptic differential equations. Poritsky discusses improvement of convergence. Weinberger obtains bounds for the square of the norm of the error in the Rayleigh-Ritz procedure, on the assumption that lower bounds for the desired proper numbers are known.

C. TRUESDELL

The Johns Hopkins University
Baltimore, Maryland

50 [V, W, Z].—BENJAMIN MITTMAN & ANDREW UNGAR, *Computer Applications—1960*, The Macmillan Co., New York, 1961, vi + 193 p., 24 cm. Price $5.75.

This small volume contains papers from the 1960 Computer Applications Symposium sponsored by the Armour Research Foundation. It is divided into two sections. Part One, "Business and Management Applications," includes the use of automatic data processing systems for handling subscription lists, information retrieval in libraries, and filling mail-orders; Part Two, "Engineering and Scientific Applications," presents applications of computers in weather forecasting, the design of optical lens systems, and electronic data communications. The wide spectrum of applications described in these papers vividly demonstrates the impressive and fast-growing uses of digital computers in management, business, engineering and scientific research. A list of titles and authors follow:

Electronic Processing of 10 Million Subscription Records—B. H. Klyce

Prediction of Program Running Time as an Aid in Computer Evaluation—T. J. Tobias

A COBOL Processor for the UNIVAC 1105—J. J. Jones

The Computer in the Library—V. W. Clapp

Computer Control of Mail-Order House Operations (IBM-650 Tape RAMAC)—S. Kritzik

An Electronic Computer in Economic Research—M. H. Schwartz

A Generalized Brokerage Accounting System (RCA 501)—A. B. Goldstein
Solution of Naval Numerical Weather Problems (CDC 1604)—P. M. Wolff
Systems and Standards Preparations for a New Computer (Philco 2000)—H. S. Bright
Computer Design of Optical Lens Systems (IBM-704)—J. C. Holladay
LOGLAN and the Machine—J. C. Brown
Data Communications Between Remote Machines—V. N. Vaughn, Jr.
Some Observations on ALGOL in Use (Burroughs 220)—J. G. Herriot
The Role of a Professional Society in Program Exchange—W. M. Carlson

MILTON SIEGEL

Applied Mathematics Laboratory
David Taylor Model Basin
Washington 7, D. C.

51 [W].—A. CHARNES & W. W. COOPER, *Management Models and Industrial Applications of Linear Programming*, Vol. II, John Wiley & Sons, Inc., New York, 1961, xiv + 861 p., 26 cm. Price $11.75.

Those familiar with the previous work of the authors will expect, and find, a well-written work thoroughly covering the field laid out for it. The emphasis is on linear programming, with some discussion and illustration by means of simplified applications to industry. Relatively little space is devoted to the practice of formulating and constructing mathematical models.

The book being reviewed is the second of two volumes. Volume I may properly be considered an introduction to Volume II. Topics covered in the second volume include the modified simplex and dual methods (with a discussion of the evolution of computer codes using the revised simplex method), transportation type models, dyadic models and subdual methods, the development of model prototypes and compression, networks and incidence type models, game theory and linear programming, and a collection of miscellaneous topics, including integer programming. Appendices treat such topics as the double-reverse method, mixing routines, and saddle points.

The treatment is broad, fully explained and amply illustrated numerically. The exposition, in this reviewer's opinion, would have been smoother if less recourse were taken to extensive footnotes. The footnotes per chapter frequently exceed 70 in number.

Equations have been attractively set into type, but all figures and tables have been reproduced directly from typewritten copy, which detracts from the over-all quality.

One further anomaly requires mention. Included in the book is an extensive bibliography numbering over 560 citations. On the other hand, the index to Volume II appears to be a skimpy afterthought. A book as potentially useful as this deserves a far better tool.

JACK MOSHMAN

C.E.I.R., Inc.
Arlington, Virginia

52 [W, Z].—ANDREAS DIERMER, *Das Wesen der automatisierten elektronischen Datenverarbeitung und ihre Bedeutung fur die Unternehmensleitung*, Walter De Gruyter & Co., Berlin, 1962, 240 p., 24 cm. Price DM 28.00.

This book belongs to the increasing number of works intended to introduce a certain group of readers to computers and their use. In one way, however, this book is rather unusual in that it is not clear, at least to this reviewer, what audience is actually being addressed. The title suggests that the work is directed towards readers in business management. In his foreword the author specifies this by stating as the aim of this book a presentation of the fundamental principles, technical aspects and use of computers in all phases of automatic business-data-processing. He qualifies this by subdividing this aim as follows: to present a thorough account of the logical and technical principles of computers in a manner understandable to business-oriented professionals; to investigate to what extent qualitatively given economic facts can be quantitized; to see which problems in the business area can be solved by the use of automatic high-speed computer systems.

With this program in mind the reader soon becomes disappointed, since the book concerns itself almost exclusively with the first point only, namely, the logical and technical fundamentals of computers. In fact, there are two parts, entitled, respectively, The Fundamentals of Automatic Electronic Data-processing, and The Methods of Automatic Electronic Data-processing as Means for the Ordering and Determination of Business Processes. The first part comprises almost 180 pages out of a total of 240. It is subdivided into three chapters which, in free translation, have the following headings: Historical Development; The Fundamentals of Electronic Data-processing; and the Interrelation of the Elements in Automatic Data-processing. Chapter 1 presents a very readable account of the history of computers and includes the European efforts in this field over the past 30 years. Chapter 2 should actually be entitled "The Building-elements of Electronic Computers." It first describes with surprising technical detail the functioning of vacuum tubes, germanium diodes, transistors, magnetic cores, etc., and then enters into a description of number systems, codes for the binary representation of decimals, etc. The chapter ends with a very general and not very enlightening philosophical discussion about the relation between mathematics and business-data-processing. Chapter 3 deals with the intercorrelation of the elements within a computer. Again the author goes into surprisingly technical details, which are rather out of place here, in view of the aim of the book and the intended readers. For example, while describing the decoding process of the instruction word, he discusses in great detail the working of a bistable multivibrator. Chapter 3 continues with a discussion of conditional-transfer instructions and fixed-point arithmetic. Floating-point representation is mentioned briefly.

The second part of the book has a very promising title, but it does not present very much substantial material. The author enters into a very general discussion of the processes fundamental to any business undertaking. Then, in equally general terms, he points to the necessity of business-data-processing, including a few words about programming problems and the Simplex method. This reviewer fails to see the aim in these discussions, which are presented in such generalities that they border on platitudes. In fact, it is here rather than in the earlier part that more specific in-

formation and greater attention to technical detail are needed in order to substantiate the numerous generalizations made.

In this reviewer's opinion, the book misses the mark and does not even meet its own aims. The largest part of the work deals with the principles of computer-building, and appears to be too technical for the business-oriented reader, while not being technical enough for the professional engineer or scientist. The rest of the book is simply too general to be useful.

WERNER C. RHEINBOLDT

University of Maryland
College Park, Maryland

53 [X].—P. G. GUEST, *Numerical Methods of Curve Fitting*, Cambridge University Press, New York, 1961, xiv + 422 p., 23 cm. Price $15.00.

This book offers a solid treatment of curve fitting or, if preferred, regression analysis. It is intended primarily for students and graduates in Physics. Chapters 1 through 4 constitute a quick course in requisite statistical inference; Chapters 5 and 6 introduce regression theory and the fitting of a straight line; Chapters 7 through 12 cover a range of topics, including polynomial regression, standard deviations of estimates, grouping of observations, special functions regression, and multiple regression.

The emphasis is certainly on the practical side, although "it is intended that the book should cover the theoretical aspects of curve fitting and full derivations of all formulae are given." For these aspects, a knowledge of calculus is assumed.

The book is quite complete in its treatment of the problems under consideration. Calculating schemes are given (primarily for a desk machine) and great care is taken in the carrying out of specific problems drawn from physics. Chapter 12 contains a guide to the more important calculating schemes for the problems considered and provides extra illustrations of commonly occurring problems. These traits should be greatly appreciated by the worker who seeks procedural method.

Some minor comments seem appropriate. Although the accomplishment of all derivations is commendable, while the stress is laid on practicality, the usual problems arise. Thus on page 3, the statement "it will often be true that the value obtained for η does not depend on the value x of ξ" launches the reader into the notion of statistical independence. More generally, it may be expected that considerable difficulty will be met in reading through the book, unless the reader has more maturity than is indicated by the background expected. For dealing with specific sections this difficulty should be considerably reduced, especially with the aid of the guide in Chapter 12.

Some notation and terminology is disturbing to a statistician. For example, this reader would prefer some stress on the linearity of estimators in the statements of Gauss-Markov theorems.

These are small criticisms of what appears to be an excellent storehouse of information for the practical curve fitter.

N. DONALD YLVISAKER

University of Washington
Seattle 5, Washington

54 [X].—R. ZURMÜHL, *Praktische Mathematik für Ingenieure und Physiker*, Springer Verlag, Berlin-Wilmersdorf, 1961, xv + 548 p., 24 cm. Price DM 29.40.

This book, the third edition of Zurmuhl's *Praktische Mathematik*, is intended to introduce the reader to many of the numerical methods for the solution of those mathematical problems that are most frequently met by the engineer. A small part of the presentation is devoted to graphical procedures for the solution of problems, but the bulk of the book deals with computational methods. The tool of computation which the author has in mind is primarily the desk computer, sometimes the slide rule, and only occasionally the electronic high-speed computer.

The book assumes as prerequisite not much more than a familiarity with the contents of a good calculus course and with the elements of linear algebra and of differential equations. The computational methods presented are described in considerable detail, including a review of much of the underlying theory.

The main topics are methods of solution for equations in one unknown (especially algebraic equations), elimination and iteration procedures for systems of linear equations, the study of matrices and their eigenvalue problems, interpolation, and numerical as well as graphical integration. The chapter on statistics is greatly enlarged in the third edition. It is followed by a chapter on the presentation of arbitrary functions, which contains a section on Fourier Analysis. The last two chapters deal in considerable detail with ordinary differential equations, the first with initial-value problems, and the other with boundary-value and eigenvalue problems. Partial differential equations are considered to be outside the scope of the book.

The presentation does not claim to be comprehensive, nor is it intended to be a mere conglomeration of methods of solution for mathematical problems. While the book does not intend to prove all theorems used, it avoids the other extreme of merely enumerating results. The reader is not only presented with the details of the numerical solution of a problem, but he is also given a thorough introduction to the ideas that lead up to the method, and so he is forewarned against applying the method to cases where its validity has not been established. While the book assumes that the reader is primarily interested in the use of a desk calculator, the greater part of the book will serve well as an introduction to methods which are of great importance in the preparation of problems for an electronic high-speed digital computer.

F. THEILHEIMER

Applied Mathematics Laboratory
David Taylor Model Basin
Washington, D. C.

TABLE ERRATA

312.—Edward S. Allen, *Six-Place Tables*, seventh edition, McGraw-Hill Book Company, New York, 1947.

In the table of common logarithms of numbers (p. 24–50) the following five corrections should be made.

number	logarithm	
	for	*read*
2443	387924	387923
4990	698100	698101
8652	937116	937117
8854	947139	947140
8884	948608	948609

On page 164, in the table of natural logarithms, the following correction is necessary in ln 6.40: *for* 1.856280, *read* 1.856298.

J. W. W.

313.—H. B. Dwight, *Tables of Integrals and other Mathematical Data*, fourth edition, The Macmillan Company, New York, 1961.

On pages 324–327 the following 72 unit corrections in the last decimal place should be made in the values of the Kelvin functions appearing in Table 1050.

Increases of a unit in the final place are required as follows: *ber* x, for $x = 3.0$, 4.4, 9.0; *bei* x, for $x = 5.7$, 6.1, 7.9, 8.5, 8.9, 9.0, 9.4; *bei'* x, for $x = 8.6$, 8.7; *ker* x, for $x = 0.2$, 2.9, 6.4, 6.5, 6.6, 8.4; *kei* x, for $x = 0.8$, 5.9, 8.8, 9.1, 10.0; *ker'* x, for $x = 0.1$, 0.3, 3.5, 7.2, 7.3, 7.4, 9.7; *kei'* x, for $x = 5.0$, 5.1, 5.2, 5.6.

Decreases of a unit in the last place are required as follows: *ber* x, for $x = 6.7$; *bei* x, for $x = 9.6$; *ber'* x, for $x = 4.0$, 7.6, 8.7; *bei'* x, for $x = 5.7$, 5.8, 6.0, 6.2, 6.3, 6.8, 7.1, 7.6; *ker* x, for $x = 3.4$, 4.1, 4.2, 4.3, 4.6, 5.9, 8.2; *kei* x, for $x = 2.8$, 3.1, 6.7, 8.3; *ker'* x, for $x = 1.6$, 3.8, 5.5, 5.6, 6.5, 7.0; *kei'* x, for $x = 3.3$, 3.5, 4.0, 4.1, 4.5, 4.7, 7.7, 8.2.

All these errors appear in the revised edition (1947) and in the third edition (1957). Additional errors appearing in this table in the first edition (1934) do not appear in the later editions.

J. W. W.

314.—Theodore E. Sterne, *An Introduction to Celestial Mechanics*, Interscience Publishers, Inc., New York, 1960.

On page 187 in formula (6.3-3) the coefficient of ∇^8 is erroneously given as 37553/518400, instead of the correct value, 33953/518400.

H. E. Salzer

Convair Astronautics Division
General Dynamics Corporation
San Diego, California

CORRIGENDUM

MARTIN GREENBERGER, "An *a priori* determination of serial correlation in computer generated random numbers," *Math. Comp.*, v. 15, 1961, p. 383–389.

The first line of equation (9) on page 385 should read

$$\sum_{x=0}^{P-1} xq = \frac{\lambda P^2}{2} - \frac{\lambda P}{2} - \frac{\mu^2}{2\lambda} - \frac{\mu}{2} - \frac{P^2}{2\lambda^2}\sum_{q=1}^{\lambda} q^2 + \frac{(\lambda + 2\mu)P}{2\lambda^2}\sum_{q=1}^{\lambda} q$$

A.
B.
C.
D.
E.
F.
G.
H.
I.
J.
K.
L.
M.
N.
O.
P.
Q.
R.
S.
T.
U.
V.
W.
X.
Z.

CLASSIFICATION OF REVIEWS

A. Arithmetical Tables, Mathematical Constants

B. Powers

C. Logarithms

D. Circular Functions

E. Hyperbolic and Exponential Functions

F. Theory of Numbers

G. Higher Algebra

H. Numerical Solution of Equations

I. Finite Differences, Interpolation

J. Summation of Series

K. Statistics

L. Higher Mathematical Functions

M. Integrals

N. Interest and Investment

O. Actuarial Science

P. Engineering

Q. Astronomy

R. Geodesy

S. Physics, Geophysics, Crystallography

T. Chemistry

U. Navigation

V. Aerodynamics, Hydrodynamics, Ballistics

W. Economics and Social Sciences

X. Numerical Analysis and Applied Mathematics

Z. Calculating Machines and Mechanical Computation

Mathematics of Computation

TABLE OF CONTENTS

JULY 1962

Some Relations and Values for the Generalized Riemann Zeta Function
ELDON R. HANSEN & MERRELL L. PATRICK 265

Euler's Constant to 1271 Places ..... DONALD E. KNUTH 275

A Calculation of the Number of Lattice Points in the Circle and Sphere
W. FRASER & C. C. GOTLIEB 282

Sign Wave Analysis in Matrix Eigenvalue Problems
K. M. BROWN & P. HENRICI 291

Acceleration Techniques for Iterated Vector and Matrix Problems
P. WYNN 301

Computing Error Bounds in Solving Linear Systems ..... J. SCHRÖDER 323

A Note on Finite Difference Methods for Solving the Eigenvalue Problems of Second-Order Differential Equations ..... M. R. OSBORNE 338

A Method for Computing the Circular Coverage Function
A. R. DIDONATO & M. P. JARNAGIN 347

Optimum-Point Formulas for Osculatory and Hyperosculatory Interpolation
HERBERT E. SALZER 356

TECHNICAL NOTES AND SHORT PAPERS

A Heuristic Asymptotic Formula Concerning the Distribution of Prime Numbers ..... PAUL T. BATEMAN & ROGER A. HORN 363

Primitive Polynomials (Mod 2) ..... E. J. WATSON 368

Note on Latent Roots and Vectors of Segments of the Hilbert Matrix
H. H. DENMAN & R. C. W. ETTINGER 370

Expansions of Jacobian Elliptic Functions in Powers of the Modulus
Y. K. LIN & F. A. LEE 372

REVIEWS AND DESCRIPTIONS OF TABLES AND BOOKS 376

KENNEDY 29, KORTUM & MCNIEL 30, SALZER & RICHARDS 31, FORSYTHE & WASOW 32, JORDAN 33, CONNOR & YOUNG 34, SMIRNOV 35, DEKANOSIDZE 36, FADDEYEVA & TERENT'EV 37, KIREEVA & KARPOV 38, LUK'IANOV, TEPLOV & AKIMOVA 39, SHERRY 40, CERRILLO & KAUTZ 41, DWIGHT 42, HUGHES & DODGE 43, BOIS 44, LEDLEY 45, BROUWER & CLEMENCE 46, ZAIDEL', PROKOF'EV & RAISKII 47, BARK, GANSON & MEISTER 48, LANGER 49, MITTMAN & UNGAR 50, CHARNES & COOPER 51, DIERMER 52, GUEST 53, ZURMÜHL 54

TABLE ERRATA 405

ALLEN 312, DWIGHT 313, STERNE 314

CORRIGENDUM 406

GREENBERGER

Published Quarterly for the

National Academy of Sciences—National Research Council

By the

American Mathematical Society